UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

TESIS DOCTORAL

Towards AI-based Network Programmability as an enabler for Zero-touch management and orchestration in B5G infrastructures

Hacia la programabilidad de red basada en IA para habilitar la gestión y orquestación Zero-touch en infraestructuras B5G

D. Jorge Gallego Madrid

2024

UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

TESIS DOCTORAL

Towards AI-based Network Programmability as an enabler for Zero-touch management and orchestration in B5G infrastructures

Hacia la programabilidad de red basada en IA para habilitar la gestión y orquestación Zero-touch en infraestructuras B5G

Autor: D. Jorge Gallego Madrid

Directores: Dr. Ramón J. Sánchez Iborra y
Dr. Antonio F. Skarmeta Gómez

UNIVERSIDAD DE
MURCIA

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD
DE LA TESIS PRESENTADA EN MODALIDAD DE COMPENDIO O ARTÍCULOS PARA OBTENER EL TÍTULO DE DOCTOR
Aprobado por la Comisión General de Doctorado el 19-10-2022

D./Dña. Jorge Gallego Madrid

doctorando del Programa de Doctorado en

Informática

de la Escuela Internacional de Doctorado de la Universidad Murcia, como autor/a de la tesis presentada para la obtención del título de Doctor y titulada:

Towards AI-based Network Programmability as an enabler for Zero-touch management and orchestration in B5G infrastructures / Hacia la programabilidad de red basada en IA para habilitar la gestión y orquestación Zero-touch en infraestructuras B5G

y dirigida por,

D./Dña. Ramón J. Sánchez Iborra

D./Dña. Antonio F. Skarmeta Gómez

DECLARO QUE:

La tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la Ley de Propiedad Intelectual (R.D. legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, modificado por la Ley 2/2019, de 1 de marzo, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita, cuando se han utilizado sus resultados o publicaciones.

Además, al haber sido autorizada como compendio de publicaciones o, tal y como prevé el artículo 29.8 del reglamento, cuenta con:

- *La aceptación por escrito de los coautores de las publicaciones de que el doctorando las presente como parte de la tesis.*
- *En su caso, la renuncia por escrito de los coautores no doctores de dichos trabajos a presentarlos como parte de otras tesis doctorales en la Universidad de Murcia o en cualquier otra universidad.*

Del mismo modo, asumo ante la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad del contenido de la tesis presentada, en caso de plagio, de conformidad con el ordenamiento jurídico vigente.

En Murcia, a 10 de junio de 2024

Fdo.: Jorge Gallego Madrid

Firmado por GALLEGO MADRID JORGE - ***7724** el día 10/06/2024 con un certificado emitido por AC FNMT Usuarios

A mis padres.

Agradecimientos

En primer lugar, quiero dar mi más sincero agradecimiento a mis directores de tesis: Antonio y Ramón. Gracias por darme la oportunidad de entrar en el mundo académico y confiar en mi desde el principio. Por iniciarme en el mundo de la investigación y hacerme descubrir qué es lo que me apasiona de verdad. Por apoyarme en todo momento, desde que todavía no era ni graduado, hasta que finalmente voy a defender mi tesis doctoral. Por empujarme a sacar lo mejor de mí y ayudarme a lanzar mi carrera investigadora. Por todos los consejos y todo vuestro tiempo. Muchas gracias por haberme ayudado a completar esta etapa, que no habría sido posible sin vuestra apuesta por mí.

Doy las gracias a mis padres, a quien les debo todo lo que fui, soy y seré. Gracias por llevarme de la mano durante todo este trayecto. Gracias por vuestro amor y vuestro cariño. Gracias por guiarme por el camino que es la vida de la mejor manera posible. Gracias por vuestro sacrificio, por tantas horas dedicadas a mí, a mi formación académica, a mi formación profesional y a mi formación como persona. Este logro es más vuestro que mío. Nunca os podré devolver todo lo que habéis dado por mí. Os quiero.

También quiero dar las gracias a toda mi familia y, especialmente, a mis abuelos y abuelas, que siempre cuidaron de mí con infinito cariño y me motivaron a conseguir lo que me propusiera. Vosotros decíais que ojalá este chiquillo llegara a ser grande en lo que decidiera hacer. Yo solo espero que os sintáis orgullosos de mis pequeños logros desde donde estéis, porque si he llegado hasta aquí, es gracias a vosotros.

Gracias a ti, Marta, por ser mi compañera inseparable en todo este camino. Por apoyarme todos y cada uno de los días. Por tener siempre una sonrisa para mí. Por escucharme. Por quererme. Por motivarme a seguir dando mi mejor versión. Por estar conmigo en los buenos momentos, pero, sobre todo, en los malos. Por seguir cumpliendo nuestras metas como siempre, juntos. Gracias.

A mis amigos, la familia que se elige, uno de mis pilares fundamentales. Cristina, Edu, Emilio, Estefanía, Gema, Jesús y José (en estricto orden alfabético). Mis amigos de la universidad, del instituto, del colegio, de Fulbright y de Estados Unidos. Y todos los demás con los que he compartido este viaje. Gracias por estar siempre ahí. Por vuestra amistad incondicional. Por vuestro cariño. Por las risas infinitas.

A mis compañeros de investigación, del departamento, de T3, de Dibulibu, de Pleiades y de más allá. Luis, Álex, Ana, Rodrigo, Jesús, Lola y Julio. Con quien tantas horas he compartido, tanto para lo bueno como para lo malo. Gracias por vuestro apoyo y vuestra amistad. Mención especial para Jordi, la luz que ilumina el camino, que tanto me ayudó en mis primeros años, y que sigue haciéndolo. Y a Jesús Sánchez, que fue el que me mostró por primera vez el significado de la palabra investigación.

Gracias a mis profesores, a todos ellos, por transmitirme vuestro conocimiento y vuestro amor por la docencia. Conseguisteis despertar en mí una pasión que nunca esperé desarrollar: dar clase en la universidad. Gracias a Rafa, con quien tantas horas de clase y de conversación he compartido.

Gracias a todos, de corazón, por ayudarme a llegar hasta aquí.

"¿Importa el destino? ¿O es el camino que emprendemos? Declaro que ningún logro tiene tan gran sustancia como el camino empleado para conseguirlo.
No somos criaturas de destinos. Es el viaje el que nos da la forma".
El camino de los reyes

"The city's central computer told you?
R2-D2, you know better than to trust a strange computer".
C-3PO

Contents

List of Figures

List of Tables

Glossary

AI	Artificial Intelligence.
API	Application Programming Interface.
ASIC	Application-Specific Integrated Circuit.
B5G	Beyond 5G.
BPF	Berkeley Packet Filter.
CAPEX	Capital Expenditure.
CNN	Convolutional Neural Network.
DL	Deep Learning.
DNN	Deep Neural Network.
DNS	Domain Name Server.
DRL	Deep Reinforcement Learning.
eBPF	extended Berkeley Packet Filter.
eMBB	enhanced Mobile Broadband.
ENI	Experiential Networked Intelligence.
ETSI	European Telecommunications Standards Institute.
FPGA	Field-Programmable Gate Array.
GAN	Generative Adversarial Networks.
GANA	Generic Autonomic Networking Architecture.
GBT	Gradient Boosted Tree.
GTP	GPRS Tunneling Protocol.
IoT	Internet of Things.
ISG	Industry Specification Group.
ITU	International Telecommunication Union.
KPI	Key Performance Indicator.
LSTM	Long Short-Term Memory.
MDP	Markov Decision Process.
MEC	Mobile Edge Computing.
MIMO	Multiple-input Multiple-output.
ML	Machine Learning.
mMTC	massive Machine Type Communications.

NEF	Network Exposure Function.
NFV	Network Function Virtualization.
NGN	Next-Generation Network.
NIC	Network Interface Card.
NN	Neural Network.
P4	Programming Protocol-independent Packet Processors.
PISA	Protocol Independent Switching Architecture.
QoE	Quality of Experience.
QoS	Quality of Service.
RAN	Radio Access Network.
RF	Random Forest.
RISC	Reduced Instruction Set Computing.
RL	Reinforced Learning.
RNN	Recurrent Neural Network.
RTT	Round Trip Time.
SDG	Sustainable Development Goal.
SDN	Software Defined Networking.
SL	Supervised Learning.
SLA	Service Level Agreement.
SON	Self-Organizing Network.
SVM	Support Vector Machine.
TCP	Transmission Control Protocol.
UPF	User Plane Function.
uRLLC	ultra Reliable Low Latency Communications.
V2I	Vehicular to Infrastructure.
V2P	Vehicular to Pedestrian.
V2V	Vehicular to Vehicle.
V2X	Vehicular to Everything.
VNF	Virtual Network Function.
VXLAN	Virtual eXtensible LAN.
XDP	eXpress Data Path.
ZSM	Zero-touch Network and Service Management.

Capítulo 1

Resumen

1.1. Motivación

La llegada de las redes de próxima generación (NGNs) traerá consigo una serie de desafiantes requisitos para las infraestructuras de comunicaciones que las soportarán. Serán necesarias una mayor flexibilidad y una orquestación automática, gracias a la integración de la inteligencia artificial (IA) en los diferentes niveles, para alcanzar anchos de banda superiores y bajas latencias, entre otros indicadores de rendimiento de interés [1, 2]. Se espera que estas redes de comunicaciones se conviertan en los cimientos de futuros servicios disruptivos para múltiples verticales, como por ejemplo, la Industria 4.0, conducción autónoma y cooperativa, servicios de telepresencia u holográficos, etc. [3]. Para lidiar con la esperable explosión de nuevas aplicaciones, numerosas soluciones para la gestión de la red están surgiendo con el objetivo de mantener bajo control la alta complejidad de la arquitectura disgregada que se espera que tengan los sistemas más allá de 5G (B5G). Esto se debe a que los servicios y las aplicaciones del futuro demandarán unos niveles de rendimiento sin precedentes de las infraestructuras de comunicaciones subyacentes, junto con nuevos métodos para la gestión de la red incapaces de ser controlados por humanos sin la ayuda de la IA. Además, para responder a todos estos requisitos y satisfacer las necesidades de los usuarios, la transición a las NGNs implica una enorme inversión en infraestructura física, así como en nuevo hardware y software actualizados.

Durante las últimas dos décadas, se han propuesto numerosos frameworks de gestión y orquestación de la red siguiendo diferentes aproximaciones. Sin embargo, aunque muchas de ellas ya se han implementado de forma exitosa en entornos de producción, todavía es necesario su evolución y mejora para poder soportar los servicios y aplicaciones B5G que vienen de camino, tal y como se ha mencionado anteriormente. En este contexto, la programabilidad y la flexibilidad de la red se posicionan como dos de los pilares fundamentales de las NGNs y deben de dirigir el desarrollo de las funciones de red futuras [4]. En esta línea, las redes definidas por software (SDN) y la virtualización de las funciones de red (NFV) han aparecido durante los últimos años como tecnologías prometedoras para habilitar el desarrollo de las infraestructuras de red del futuro. Aunque su máximo potencial está lejos de ser alcanzado. A pesar de que estas tecnologías han proporcionado nuevas dimensiones a la flexibilidad y programabilidad de la red, la softwarización de la red exige soluciones más innovadoras. El objetivo es alcanzar una gestión y orquestación de los recursos de red autónoma y con alto rendimiento para poder responder a las necesidades de los usuarios en términos de la calidad de experiencia (QoE). Además, el panorama se está volviendo más complejo con la aparición continua de redes multi-operador que comparten de

forma virtual la misma infraestructura física, tal y como se propone en el paradigma de 6G [4]. Bajo esta visión, en la nueva era de las comunicaciones, habrá billones de humanos, vehículos y pequeños dispositivos interconectados y generando tremendas cantidades de datos a través de infraestructuras de red heterogéneas que serán gestionadas por distintos actores: empresas de telecomunicaciones, operadores de infraestructura, proveedores de servicios, etc.

Las NGNs deberían liderar en los próximos años el camino para conseguir una computación distribuida, autónoma, sostenible, flexible y confiable. Desde el punto de vista de la gobernanza, la IA es la pieza clave para conseguir un mantenimiento efectivo de los complejos servicios que operen en las redes del futuro. La softwarización de la red acaba de empezar con la proliferación de una miríada de tecnologías de virtualización, y su interoperabilidad supone un desafío crucial para la operación de las redes. En consecuencia, las funciones de gestión y orquestación potenciadas con IA serán imprescindibles para automatizar el proceso de toma de decisiones. Machine Learning (ML) es la tecnología que se está adoptando para proveer de inteligencia a estos sistemas. Estas técnicas dotarán a la infraestructura de red con capacidades de gobierno autónomas para hacerla auto-adaptable considerando en tiempo real sus propias necesidades y las de los usuarios o servicios. La evolución de las funciones de red integradas con la IA permitirá la orquestación predictiva para optimizar aún más la gestión del tráfico, la localización de los recursos y la configuración de los servicios basándose en las necesidades esperadas y anticipándose a las demandas de los usuarios. Es esto último lo que permitirá conseguir la gestión de servicios y de red zero-touch (ZSM), donde la intervención humana será reducida a cero y la red empezará a operar de forma autónoma y, por tanto, su eficiencia será optimizada a niveles que nunca se han visto.

También se espera que las arquitecturas de red y de computación tradicionales, que se encuentran ahora mismo evolucionando hacia el paradigma de computación en el Edge (MEC), converjan en un único continuo de computación que cubra desde la nube hasta los usuarios o dispositivos finales (cloud-core-edge-dispositivos). Como resultado, la computación ubicua democratizará el uso de los recursos entre todos los actores haciendo uso de la red, provocando que la capacidad de respuesta se decremente significativamente. Además, el volumen de tráfico del modelo clásico cliente-servidor será reducido, reduciendo la latencia de la red. Gracias a la tendencia de virtualizar los servicios como cadenas de funciones de red, será más sencillo que nunca distribuir las tareas de computación y networking por todo el continuo. La importancia de este paradigma es incluso mayor si consideramos el objetivo de las NGNs de alcanzar latencia cero, lo cual es posible únicamente con la introducción de procesamiento local o muy cercano al usuario.

Toda esta revolución trae consigo una inversión considerable en equipamiento, por lo que reducir los costes es fundamental. Además, es crucial para nuestro futuro mejorar la eficiencia energética y reducir la huella ecológica. Una de las formas de abordar estos problemas es explotando y reutilizando hardware de bajo coste e introduciendo virtualización ligera y gestión de los recursos con una granularidad muy baja, mientras se mantienen niveles altos de rendimiento en la computación y networking. Esto implica la adopción de nuevos modelos de virtualización y de composición de servicios para aprovechar los nuevos paradigmas de gestión y orquestación. Además, se abre la puerta a que dispositivos con recursos limitados y no especializados participen en las operaciones de la infraestructura de red. El anteriormente mencionado continuo de computación y networking beneficiará a este enfoque, ya que tanto las funcionalidades del plano de control como la programabilidad del plano de datos se gestionarán de forma integrada a lo largo de este continuo extremo a extremo. Además, las velocidades de tráfico que se espera alcanzar en las NGN exigirán capacidades de procesamiento de paquetes de altísima eficiencia al hardware para acercar la inteligencia lo máximo posible a los usuarios y reducir la latencia al mínimo. Para ello, las técnicas de reprogramabilidad del plano de datos tendrán que formar parte de esta transformación para convertirse en los cimientos de la arquitectura de red.

Uno de los mercados verticales en los que más repercutirá esta evolución del paradigma es el de los vehículos. En los últimos años, las comunicaciones vehículo a todo (V2X) han despertado un gran interés, tanto en la industria como en el mundo académico. Las mejoras esperadas de la seguridad vial, la disminución de la huella de carbono, la introducción de sistemas de visualización holográfica, el infoentretenimiento inmersivo y la mejora de la experiencia del usuario exigen un cambio de la forma en

que se conciben las redes vehiculares. V2X prevee muchos actores implicados en un entorno dinámico, a la hora de considerar una amplia gama de estrategias de conectividad para cubrir todos los escenarios, por ejemplo, vehículo a vehículo (V2V), vehículo a peatón (V2P) y vehículo a infrastructura (V2I). [5]. Para proporcionar eficazmente una calidad de experiencia adecuada a todas las partes interesadas, V2X llevará al límite las capacidades de las arquitecturas de red existentes. Por este motivo, necesita como base una nueva generación de infraestructuras de red que se apoyen en el paradigma ZSM. Además, debido a la extrema heterogeneidad presente en el ecosistema de la automoción, en términos de escenarios de comunicación, rigurosos requisitos y capacidades informáticas limitadas, se necesitan nuevas soluciones que permitan una toma de decisiones autónoma y flexible al tiempo que se mantiene un alto rendimiento.

Si bien es cierto que la comunidad investigadora ha investigado en profundidad estos temas y ha proporcionado avances significativos en los retos identificados durante los últimos años [6], aún quedan algunos puntos por conectar y múltiples huecos en el estado del arte. Estas fueron las semillas que motivaron el inicio de este viaje de doctorado. En el punto de partida de la tesis, las técnicas de ML empezaban a utilizarse para enriquecer las arquitecturas de red con inteligencia y gestión autónoma, aunque se encontraban en una fase inicial y, por tanto, aún tenían algunas carencias. Su aplicación se estudiaba como una solución limitada para manejar sólo ciertos segmentos de la red mediante el uso de modelos muy específicos. Sin embargo, era necesario ampliar esta visión restringida para explorar la gestión holística y automática de infraestructuras de red complejas. Como propone el paradigma ZSM, esto sólo puede lograrse analizando el problema desde dos perspectivas diferentes, aunque alineadas: la gestión del plano de datos y la administración del plano de control. La primera requiere tecnologías rápidas y eficientes para procesar la cantidad ingente de tráfico que atravesará la infraestructura de red en las NGNs. De este modo, se echaban en falta estudios que exploraran y validaran estas tecnologías de gestión de tráfico en entornos reales para seleccionar las mejores soluciones para cada escenario, teniendo en cuenta el continuo único de computación y networking que abarca desde los dispositivos finales hasta la nube. Del mismo modo, la bibliografía apenas aborda la integración entre las técnicas de reprogramabilidad del plano de datos y los modelos de gestión inteligente del tráfico, lo que pone de manifiesto la necesidad de gestionar automáticamente las funcionalidades del plano de control y el procesamiento del plano de datos de forma transparente. Además, la validación de estas soluciones solía hacerse considerando la disponibilidad de altas capacidades de recursos de computación, dejando atrás una amplia gama de dispositivos más limitados, situados en el borde de la red. Sobre el diseño de las infraestructuras del futuro, los enfoques generalistas eran la tendencia para cubrir las múltiples dimensiones relacionadas con la evolución de las redes de comunicaciones. Se veía necesaria una mayor concreción para extraer métricas tangibles y objetivas que debían cumplir las infraestructuras. Por último, también era necesario diseñar arquitecturas de gestión que permitieran gestionar las aplicaciones y servicios futuros de forma eficiente, aprovechando al máximo la infraestructura de red subyacente y sus recursos.

La presente tesis doctoral describe los resultados de la investigación sobre el diseño, implementación y validación de soluciones de reprogramabilidad de red basadas en ML, como facilitadoras de la adopción del paradigma ZSM en infraestructuras de red de próxima generación. Las propuestas de esta tesis permiten la integración de los modelos de IA en los esquemas de gestión del plano de datos en hardware genérico, con el fin de satisfacer las necesidades del futuro continuo de computación y networking. La reprogramabilidad del plano de datos se consigue gracias a eBPF [1]. Esta tecnología permite la ejecución de programas desarrollados por el usuario en el núcleo de Linux, lo que agiliza el procesamiento de paquetes y reduce al mínimo las perturbaciones en los flujos de tráfico. Además, también permite la integración de complejos algoritmos que tradicionalmente se ejecutaban en servidores dedicados en hardware genérico. Aunque la gama de algoritmos de ML es amplia y se podrían haber considerado varios, se han adoptado las redes neuronales, ya que proporcionan gran flexibilidad y alta precisión para analizar el tráfico y detectar anomalías [7]. Con estas potentes herramientas, esta tesis explora dos escenarios oportunos y diferentes, a saber, (i) el procesamiento de tráfico de gran volumen en una red

[1] https://ebpf.io/

5G con el objetivo de mejorar la calidad de servicio, y (ii) la integración de análisis de tráfico avanzado en dispositivos restringidos con foco en la ciberseguridad. Como colofón, todos los conocimientos adquiridos se ponen en práctica para idear cómo las aplicaciones y los servicios futuros dirigirán el diseño y el desarrollo de las arquitecturas NGN. Este ejercicio se canaliza a través del vertical vehicular, que ejemplifica las altas demandas de recursos que deben abordar las infraestructuras de red del futuro. Como resultado, se identifican los puntos críticos para proporcionar métricas objetivas y medibles que sirvan como indicadores de rendimiento de referencia para las infraestructuras de red. Por último, se presenta una plataforma holística destinada a establecer los componentes básicos que darán soporte a las necesidades futuras de las NGNs.

Esta tesis doctoral ha sido llevada a cabo con el apoyo de la Fundación Séneca—Agencia de Ciencia y Tecnología de la Región de Murcia (España)[2]—bajo la beca FPI *21429/FPI/20*, y cofundada por Odin Solutions S.L.[3], Región de Murcia (Spain); y por la Comisión Fulbright en España[4] bajo la beca Fulbright *00003/FLB/21*.

1.2. Objetivos y Metodología

Para avanzar en el estado del arte y hacer frente a los retos discutidos anteriormente, el objetivo principal de esta tesis es explorar las necesidades actuales de las NGNs para adoptar capacidades ZSM plenamente funcionales. Se pretende proporcionar la integración de diferentes tecnologías que permitan la gestión rápida del plano de datos, por ejemplo, el procesamiento eficiente de paquetes, dirigido por funciones de red basadas en IA que permitan la gestión autónoma de la red en tiempo real. La metodología seguida se ha centrado en abordar la convergencia de las tecnologías de programabilidad del plano de datos y los esquemas de IA. Por tanto, en primer lugar, esta tesis se centra en un meticuloso estudio para identificar algoritmos de ML que estén bien posicionados para ser adoptados por sistemas ZSM. A continuación, traslada el foco a la búsqueda, selección y evaluación del rendimiento de tecnologías que permitan un procesamiento rápido de paquetes con un impacto casi nulo en las prestaciones de la red. Después, ambas partes se fusionan para lograr la reprogramabilidad del plano de datos dirigida por modelos inteligentes que permiten la creación de herramientas muy adecuadas para infraestructuras del futuro basadas en ZSM. Su viabilidad se demuestra mediante la validación de su ejecución en hardware genérico, mostrando la integración de procesos de computación tradicionalmente exigentes en recursos en equipos ligeros. Finalmente, el conocimiento y experiencia adquiridos en las etapas anteriores se explotan para vislumbrar aspectos clave en el diseño de infraestructuras 6G en el contexto del vertical vehicular, que plantea numerosos retos en el diseño de las arquitecturas de red para satisfacer los estrictos requisitos que supone el desarrollo de servicios V2X. Para alcanzar la meta principal de la investigación de esta tesis, se identificaron los siguientes objetivos:

- **Objetivo 1**: analizar y estudiar el estado del arte en la aplicación de técnicas basadas en ML para tareas de gestión y orquestación de redes.
- **Objetivo 2**: identificar algoritmos de ML adecuados y eficientes para ser adoptados por sistemas ZSM en infraestructuras de red complejas.
- **Objetivo 3**: seleccionar tecnologías de programabilidad de red adecuadas para abordar los retos de las redes del futuro.
- **Objetivo 4**: implementar e integrar técnicas de procesamiento de tráfico basadas en ML en infraestructuras B5G para proporcionar una gestión de paquetes autónoma y rápida sin penalizar el rendimiento.

[2] https://fseneca.es/
[3] https://odins.es/
[4] https://fulbright.es/

- **Objetivo 5**: evaluar las soluciones desarrolladas en hardware genérico para permitir su despliegue en todo el continuo de computación y networking, desde los dispositivos finales hasta la nube.
- **Objetivo 6**: capitalizar la experiencia adquirida para vislumbrar los puntos clave que dirigirán el diseño de las infraestructuras 6G.
- **Objetivo 7**: explorar cómo los futuros servicios y aplicaciones con altas demandas de recursos de red y computación desafiarán a las redes 6G en el contexto de un vertical tan exigente como el vehicular.

Para alcanzar dichos objetivos, el trabajo realizado durante este doctorado se dividió en diferentes líneas de investigación correspondientes a cada uno de los objetivos. Una vez completadas todas ellas, convergieron para componer esta tesis doctoral. En primer lugar, hubo una fase de búsqueda y estudio, en la que se revisó la literatura para conocer en profundidad el estado del arte de ZSM, ML aplicado a redes de comunicaciones, actividades de estandarización y proyectos de investigación enfocados en NGNs. Una vez hecho esto, se exploraron y examinaron las herramientas y tecnologías más utilizadas en estas áreas, con el fin de seleccionar las más adecuadas para lograr la reprogramabilidad inteligente de la red en escenarios ZSM. eBPF fue la tecnología de procesamiento de paquetes seleccionada tras compararla experimentalmente con P4. Se realizó una extensa evaluación de rendimiento para obtener información sobre su capacidad para procesar tráfico a alta velocidad, en términos de ancho de banda máximo permitido, latencia y uso de CPU; evidenciando que eBPF es una solución flexible, sencilla y rentable que puede desplegarse en múltiples escenarios sin requisitos de hardware prohibitivos. Al utilizar eBPF como tecnología de procesamiento de tráfico, la siguiente etapa consistió en seleccionar los algoritmos de ML adecuados para analizar el tráfico y reaccionar automáticamente ante cambios, interrupciones o amenazas en la red. La revisión de la literatura inicial reveló que las redes neuronales son uno de los pilares fundamentales para permitir la toma de decisiones inteligentes en entornos ZSM. De este modo, se realizó un desarrollo para combinar el procesamiento de paquetes en eBPF con redes neuronales para obtener capacidades de gestión y control de la red habilitadas por ML. Como resultado, se integró en el kernel de Linux un pipeline de manejo de paquetes basado ML, acelerando así la ejecución eficiente de algoritmos de IA. Para validar la implementación desarrollada, se exploró un caso de uso real, en el que la solución procesaba el tráfico en vivo mientras se utilizaba una red neuronal para detectar ataques en una red IoT. Finalmente, con todo el conocimiento adquirido durante el desarrollo de las líneas de investigación mencionadas, se realizó un ejercicio final para concebir los aspectos clave de networking y computación que impulsarán el diseño de las futuras infraestructuras 6G. El exigente y desafiante vertical vehicular fue estudiado desde el punto de vista de sus futuras aplicaciones y servicios para explorar cómo estas darán forma a la definición y despliegues de redes 6G.

1.3. Resultados

El logro de cada uno de los objetivos propuestos en esta tesis doctoral dio lugar a varias publicaciones científicas. Cuatro de estas publicaciones han sido seleccionadas para formar este compendio de tesis doctoral. Cabe destacar que estos artículos han sido publicados en revistas internacionales clasificadas en el primer cuartil del JCR de Clarivate Analytics [5]. Los principales resultados obtenidos durante esta tesis doctoral, como fruto de la investigación realizada, se presentan en la Tabla 1.1, en la que también se indican los objetivos y publicaciones asociadas.

Además de los cuatro artículos de revista que componen esta tesis, también se han publicado otros trabajos relacionados: dos capítulos de libro, diez artículos de conferencia y cuatro artículos adicionales indexados en JCR. Nótese que esta tesis se ha presentado como un compendio de cuatro artículos de investigación que representan el núcleo de este doctorado. Por lo tanto, la principal contribución de la investigación llevada a cabo está contenida en estos cuatro artículos y el resto de publicaciones amplían las aportaciones de esta tesis al estado del arte.

[5] https://jcr.clarivate.com/jcr

Tabla 1.1: Principales resultados de la tesis

Resultado	Objetivos	Publicaciones
R1. Análisis del estado del arte de los algoritmos y técnicas de ML candidatos a ser adoptados en sistemas ZSM.	1, 2, 4	[115], [117]
R2. Análisis de tecnologías de vanguardia para el procesado rápido de paquetes en redes B5G.	3, 4, 5	[116], [117]
R3. Diseño, implementación y validación de una solución basada en eBPF y P4 para proveer calidad de servicio en redes 5G multioperador.	2, 3	[116]
R4. Evaluación del rendimiento de eBPF y P4 en términos de eficiencia de procesado de paquetes y programabilidad.	3, 5	[116], [117]
R5. Diseño e implementación de un pipeline basado en ML para el procesamiento eficiente de paquetes en el kernel de Linux.	1, 2, 4, 5	[115], [117]
R6. Integración y validación en hardware genérico de una solución basada en eBPF que usa ML para reaccionar dinámicamente a ciberataques en un entorno IoT.	4, 5	[115], [116], [117]
R7. Identificación de KPIs objetivos y concretos que puedan medir el rendimiento de infraestructuras 6G.	6, 7	[118]
R8. Diseño de una plataforma 6G orquestada por ZSM capaz de gestionar el ciclo de vida completo de servicios multidominio del futuro.	1, 6, 7	[118]

A continuación, se encuentra un resumen detallado del trabajo realizado en cada una de las publicaciones que componen esta tesis doctoral, relacionándolas de forma más explícita con los resultados obtenidos. Además, los documentos completos de estas publicaciones se incluyen en el Capítulo 4.

1.3.1. Machine learning-based zero-touch network and service management: a survey

El primer trabajo que forma el compendio [115] explora técnicas de ML que puedan ser adoptadas para potenciar el paradigma ZSM. A través de una revisión en profundidad de la taxonomía de algoritmos de ML, se recapitulan y analizan múltiples propuestas encontradas en la literatura para gestionar y orquestar infraestructuras NGN. Como resultado de este estudio, se identifican las diferentes funciones de control de red que permiten la gestión automática de la misma, a la vez que se proporcionan las técnicas más adecuadas para la implementación de cada componente (**R1**). Los algoritmos de ML se suelen clasificar en cuatro familias diferentes: supervisados (SL), no supervisados (UL), semisupervisados (SSL), y de refuerzo (RL). La selección del algoritmo más adecuado es multifactorial, ya que depende de las características de la infraestructura de comunicaciones subyacente, los requisitos de tráfico de la red, el comportamiento esperado de los usuarios y el tipo de servicios y aplicaciones que operan en la red. La discusión de las propuestas que aplican IA a la gestión y orquestación de la red se clasifica según los roles de las funciones de red estudiadas dentro de la arquitectura de red: (i) inspección de flujos de paquetes, (ii) gestión multidominio, (iii) gestión de la radio, y (iv) gestión de recursos de red. En la revisión bibliográfica realizada se ha puesto de manifiesto que las técnicas SL son las más utilizadas en escenarios de inspección de flujos, ya que los operadores suelen tener sus propias políticas para clasificar el tráfico que viaja por su infraestructura, y el uso de algoritmos supervisados puede aprovechar esta información y conseguir una alta eficiencia. En escenarios multidominio, los enfoques jerárquicos funcionan mejor que las soluciones centralizadas debido a la visión parcial de la red que presentan los distintos componentes. Los algoritmos de ML típicos tienen dificultades para operar eficientemente en estos entornos complejos y cambiantes, ya que el entrenamiento en tiempo real es

muy costoso debido a la cantidad de muestras necesarias. En consecuencia, las soluciones de RL son la opción más común para recuperar información de los múltiples dominios y seleccionar el algoritmo más adecuado para predecir el comportamiento esperado. Mantener un funcionamiento en tiempo real de los algoritmos de ML en la gestión de la radio implica el uso de una enorme cantidad de recursos computacionales, lo que lo hace muy costoso debido a sus características dinámicas inherentes. De esta forma, aunque las soluciones existentes para este segmento de la red se basan en RL para adaptarse rápidamente a los cambios en el medio radioeléctrico, la mayoría de ellas realizan un entrenamiento offline que posteriormente se despliega en el entorno de producción. La gestión de recursos es un campo amplio y abarca diferentes aspectos en el control de la red, por lo que no existe una forma unificada de abordar su administración. En la literatura, predominan las soluciones SL y RL. La primera aprovecha su capacidad para hacer frente a la alta complejidad y heterogeneidad de la red, mientras que la segunda proporciona una reacción flexible y en tiempo real a los cambios en el tráfico y el comportamiento de los usuarios. Además de este análisis, este trabajo también explora las actividades de estandarización y los proyectos de investigación que trabajan en la definición y evolución del paradigma ZSM. El interés que este está despertando tanto en el mundo académico como en la industria queda patente en el gran número de iniciativas que aportan soluciones conceptuales e implementaciones reales, lo que también se analiza en profundidad. Finalmente, se discuten los retos y líneas de futuro de la gestión y orquestación automatizada de las NGNs. De esta forma, el trabajo realizado en este artículo de revista proporciona un conocimiento profundo sobre el estado actual y las tendencias futuras en sistemas ZSM basados en ML y sienta las bases para los próximos trabajos publicados que conforman el compendio.

1.3.2. Fast traffic processing in multi-tenant 5G environments: A comparative performance evaluation of P4 and eBPF technologies

El segundo trabajo del compendio [116] presenta una revisión y una evaluación de rendimiento de dos de las tecnologías más prometedoras para manejar tráfico a un alto ancho de banda, garantizando al mismo tiempo una baja latencia en el procesamiento de paquetes en infraestructuras NGN: P4 y eBPF (**R2**). P4 es un lenguaje de programación orientado a arquitecturas hardware programables (PISA), que es el modelo de plano de datos de referencia para dispositivos de red programables. La base de esta arquitectura es un pipeline donde, en primer lugar, los bits de los paquetes se alinean con los protocolos; después, múltiples acciones pueden modificar el paquete si coincide con una serie de reglas; y, por último, el flujo de bits se convierte de nuevo en paquetes antes de enviarlo de vuelta a la red. Por otro lado, eBPF es una tecnología que permite ejecutar programas en un entorno controlado y seguro dentro del kernel de Linux. Permite cargar nuevo software dentro del kernel sobre la marcha y sin modificar el código original, ampliando las capacidades del kernel de forma segura. Cuando se combina con el XDP de Linux, permite la implementación eficiente de funciones de red potentes, a la vez que ligeras y portátiles. El sistema operativo es el que garantiza la seguridad y ejecución de los programas con la ayuda de un motor de verificación y un compilador just-in-time. Ambas tecnologías se estudian en este trabajo, demostrando ser alternativas capaces de permitir el procesamiento rápido de paquetes y la programabilidad en las infraestructuras de red. En los experimentos realizados sobre una infraestructura 5G real, P4 muestra un rendimiento muy alto debido a su procesamiento de tráfico basado en hardware, teniendo como principal inconveniente su elevado coste de despliegue. Por otro lado, eBPF aún estaba en sus etapas iniciales cuando se utilizó en este trabajo para la reprogramabilidad del plano de datos, pero muestra prometedoras capacidades como un habilitador de funciones de red simple, portátil y eficiente. A pesar del brillante futuro de estas tecnologías, antes de este artículo no existían estudios en la literatura que evaluaran y compararan su rendimiento en NGNs. Por lo tanto, el propósito de este trabajo es servir como indicador de referencia del rendimiento de P4 y eBPF para el diseño, desarrollo y despliegue de funciones de red en infraestructuras ZSM. En primer lugar, se realiza una discusión sobre las ventajas e inconvenientes de ambos, junto con una revisión bibliográfica de los trabajos más destacados en la materia. De este modo, se identifican los escenarios en los que una u otra es más adecuada. A continuación, se presenta una herramienta de gestión de paquetes 5G orientada a calidad de servicio que utiliza ambas tecnologías con el objetivo de proporcionar una función de red

capaz de realizar un procesamiento temprano de paquetes en un nodo edge (**R3**). Su rendimiento se evalúa desde distintas perspectivas en un banco de pruebas 5G, por lo que se muestran resultados de distinta naturaleza. P4 obtiene las mejores cifras en cuanto a tasa de procesamiento de paquetes, lo cual no es una sorpresa; como solución basada en hardware se espera que supere a cualquier otra basada en software. Sin embargo, la tupla formada por eBPF y XDP obtiene resultados similares en escenarios de gran ancho de banda donde la longitud media de los paquetes es elevada. Además, para evaluar la capacidad máxima de estas tecnologías, los experimentos realizados utilizan volúmenes de tráfico extremos. Sin embargo, con cargas de tráfico moderadas y más típicas, la solución eBPF/XDP iguala las prestaciones de la implementada utilizando P4. A la luz de estos resultados, existen algunos escenarios en los que los requisitos de la red no son tan estrictos como en el núcleo de la red, como los casos de uso en el edge o los escenarios IoT. En estos entornos, las funciones de red ligeras y flexibles pueden colocarse fácilmente bajo demanda en distintos puntos de la red. También pueden ser útiles en sistemas ZSM, donde las condiciones de la red pueden cambiar en cualquier punto de la infraestructura y estas pequeñas funciones de red pueden ser fácilmente desplegadas con cualquier comportamiento deseado en cuestión de segundos. Además, si tenemos en cuenta el coste de la inversión, el equipamiento de red especializado es caro de desplegar en diferentes puntos de la infraestructura. Sin embargo, el despliegue de hardware básico capaz de albergar funciones de red basadas en eBPF/XDP puede reducir significativamente el coste de la infraestructura, manteniendo al mismo tiempo la capacidad de procesar tráfico de forma flexible y eficiente. Tras completar este estudio y la evaluación de prestaciones de ambas tecnologías (**R4**), eBPF se erige como una tecnología prometedora para permitir el uso de funciones de red complejas, como las que permiten los algoritmos de ML, en escenarios ZSM. De esta forma, eBPF fue la tecnología seleccionada para continuar el camino de investigación seguido en esta tesis. El siguiente paso consistió en hacer converger los conocimientos y antecedentes adquiridos con los dos primeros trabajos de esta tesis. Concretamente, el foco se centró en desarrollar funciones de red eficientes aprovechando las sinergias entre eBPF y ML para realizar tareas de red automatizadas en cualquier punto de la infraestructura sin esfuerzo, que es el reto que se aborda en el siguiente artículo de este compendio.

1.3.3. Machine learning-powered traffic processing in commodity hardware with eBPF

El tercer artículo del compendio [117] presenta el desarrollo y evaluación de una función de red que aprovecha eBPF y ML para proporcionar procesamiento inteligente dentro del kernel de Linux en hardware genérico (**R6**). La implementación permite la ejecución de tareas computacionales pesadas, como redes neuronales, de forma sencilla y flexible. Como se comenta en el segundo artículo del compendio [116], eBPF se plantea como un habilitador para desarrollar una gestión y monitorización eficiente de la red en cualquier punto de la infraestructura (**R4**), lo cual es muy valioso para desarrollar funciones de red de seguridad. Esto se debe a que la inspección del tráfico puede ser realizada por cualquier dispositivo Linux a altas velocidades con un impacto mínimo en la latencia. Por lo tanto, los accesos y ataques no autorizados pueden detectarse en tiempo real. El enfoque típico encontrado en la literatura combina el uso de algoritmos de ML que se ejecutan en el espacio de usuario y que se alimentan con los datos recogidos por sondas eBPF. Sin embargo, este desacoplamiento puede dar lugar a una degradación del rendimiento, ya que el espacio de usuario es menos eficiente para realizar tareas computacionales en comparación con una implementación en el núcleo. De este modo, la función de red desarrollada combina la inspección rápida de paquetes y la toma de decisiones basada en ML dentro del núcleo de Linux, ahorrando recursos computacionales del dispositivo y reduciendo las latencias de procesamiento. Siguiendo este enfoque, es posible incrementar el rendimiento de las tareas de red inteligente de forma flexible y portable, lo cual es de suma importancia en las NGNs. La solución implementada se valida en un caso de uso IoT en el que el uso de dispositivos con recursos computacionales limitados es la norma. Las capacidades reducidas de estos dispositivos en términos de memoria, capacidad de procesamiento y consumo de energía plantean un reto para la incorporación de funciones de redes inteligentes o ciberseguridad. En este entorno, cada nodo es un posible vector

de entrada a toda la red, por lo que reforzar sus capacidades de defensa frente a ciberataques es crítico para la robustez de la infraestructura. Gracias a la eficiencia para ejecutar código a nivel del núcleo Linux, sofisticadas soluciones de seguridad que serían pesadas de ejecutar en un contexto típico, pueden integrarse completamente en el sistema operativo, lo que permite ejecutar una nueva gama de funciones de red de seguridad en hardware genérico. Además, la inclusión de algoritmos inteligentes a este nivel también permite la detección automática y la reacción ante las amenazas a la seguridad en entornos IoT de forma flexible y ligera. Concretamente, en el caso de uso considerado, la función de red eBPF desarrollada se valida en una red 6LoWPAN que utiliza RPL. 6LoWPAN es un estándar definido por el IETF para integrar IPv6 en redes inalámbricas de baja potencia, como las definidas por el estándar IEEE 802.15.4 [8]. RPL es un protocolo de encaminamiento para este tipo de redes basado en vectores distancia que opera sobre 6LoWPAN. Es uno de los protocolos de encaminamiento IoT más utilizados debido a su facilidad para crear y compartir rutas y adaptarse a los cambios de topología. De esta forma, el escenario evaluado está formado por nodos 6LoWPAN amenazados por el ataque *Hello Flood* al protocolo RPL. En este caso, el ataque consiste en que varios dispositivos maliciosos envían peticiones de información de encaminamiento al resto de nodos, obligándoles a responder y malgastando sus recursos. El primer paso para evaluar el rendimiento de la función de red desarrollada es generar un conjunto de datos para entrenar los modelos de ML. Para ello, se despliega una red IoT utilizando el simulador de Cooja[6], una herramienta que permite la emulación de dispositivos virtuales que implementan 6LoWPAN y RPL. En este entorno se simula el ataque y se obtienen tres conjuntos de datos. A continuación, aunque múltiples algoritmos de ML podrían haber encajado en este caso de uso, el uso de redes neuronales fue la elección final, debido a su buen rendimiento y eficiencia para inspeccionar el tráfico y detectar anomalías [7]. Sin embargo, son modelos complejos y no pueden implementarse directamente en eBPF debido a las restricciones impuestas por el verificador. Por lo tanto, el punto de partida fue utilizar la implementación de red neuronal de Scikit-Learn de Python[7]. El modelo se entrenó con una división 80/20 y, tras múltiples pruebas, se obtuvo una red neuronal conformada por dos capas ocultas (compuestas por tres y dos perceptrones, respectivamente) (**R5**). El modelo alcanzó una precisión entre 0,9 y 1 en todos los escenarios considerados, aunque el foco del trabajo no es alcanzar una gran precisión en la detección de ataques sino ejecutar el modelo dentro del kernel Linux, algo no logrado hasta este punto en la literatura. Con el modelo listo, se utilizó la librería TinyML *emlearn* [9] para convertir el código Python en código C, que fue la base del programa eBPF. A continuación, este código se modificó y personalizó para ser validado por el verificador eBPF y embebido en el kernel de Linux. También se desarrolló otra versión de la función de red para desacoplar su funcionalidad, es decir, situar el análisis de paquetes a nivel de kernel, y la ejecución del modelo para su inspección en el espacio de usuario, a efectos comparativos. La evaluación del rendimiento realizada en un dispositivo real muestra una clara mejora de la solución completa en el núcleo, mejorando en un 6 % el uso de la CPU y reduciendo en un 97 % el tiempo necesario para ejecutar la red neuronal. Este resultado demuestra las ventajas derivadas de las sinergias entre eBPF y ML en el desarrollo de funciones de red complejas e inteligentes capaces de ejecutarse en hardware genérico en cualquier punto del continuo de networking y computación. Estos hallazgos también allanan el camino para la exploración del próximo paradigma en infraestructuras de comunicación, 6G, que es el tema principal del siguiente artículo del compendio de esta tesis doctoral.

1.3.4. The role of vehicular applications in the design of future 6G infrastructures

Finalmente, el cuarto y último artículo del compendio [118] tiene como objetivo aplicar todo el conocimiento adquirido durante el doctorado sobre NGNs a la próxima generación de infraestructuras de comunicaciones celulares, centrándose en el vertical vehicular. Este trabajo aborda V2X para 6G desde un enfoque centrado en los servicios, proporcionando una visión global del próximo ecosistema vehicular y su integración con 6G. El diseño de la arquitectura 5G se realizó con un enfoque centrado

[6] https://github.com/contiki-os/contiki/wiki/An-Introduction-to-Cooja
[7] https://github.com/scikit-learn/scikit-learn

en la red, que inicialmente carecía de requisitos concretos de las aplicaciones cubiertas por las tres familias de servicios, es decir, uRLLC, mMTC y eMBB. En las próximas definiciones de 6G, resulta de vital importancia considerar desde el principio una visión centrada en los servicios para alinear el diseño de las futuras infraestructuras con las demandas de las aplicaciones que operan sobre ellas. Aunque 6G pretende cubrir un amplio espectro de sectores verticales, es necesario estudiar específicamente todos ellos, huyendo de planteamientos genéricos. Así, las aplicaciones vehiculares son un buen ejemplo de un importante vertical que demandará requisitos específicos y exigentes a las futuras infraestructuras 6G. Los casos de uso característicos de este vertical demandan baja latencia y gran ancho de banda en escenarios críticos, así como robustez ante desconexiones debido a su inherente movilidad [10]. En esta línea, es crucial identificar los requisitos concretos de estos servicios para definir cómo se construirán las infraestructuras. A partir del estudio realizado revisando tanto las aplicaciones 5G en evolución hacia 6G, como los servicios disruptivos previstos que llegarán en el futuro, se identifican los requisitos de rendimiento que plantearán a la próxima infraestructura 6G, que son los siguientes: (i) calidades de servicio deterministas y anticipadas de extremo a extremo, (ii) diferentes tecnologías de acceso, (iii) orquestación y reprogramabilidad de la red basad en IA, (iv) gestión del ciclo de vida de las aplicaciones basada en IA, (v) entornos seguros y confiables de despliegue y validación de aplicaciones multidominio, (vi) percepción multidimensional y posicionamiento preciso a alta velocidad, (vii) almacenamiento en caché y frescura de los datos, (viii) privacidad y seguridad de los datos, y (ix) eficiencia energética. Todos ellos sirven de base para la propuesta específica y medible de KPIs que puedan capturar el rendimiento de la infraestructura de red (**R7**). Además, basándose en el análisis previo, también se presenta el diseño de una plataforma 6G basada en ZSM para gestionar todo el ciclo de vida de servicios y aplicaciones V2X multidominio (**R8**). La plataforma se basa en una perspectiva experimental, donde el diseño y desarrollo de futuros servicios de red debe pasar automáticamente por una etapa de validación en infraestructuras de prueba antes de su despliegue en escenarios de producción. La plataforma propuesta permite la validación de las aplicaciones en cumplimiento de ciertos requisitos que provienen de KPIs específicos. Además, el diseño se alinea con sistemas ZSM basados en IA con el objetivo de manejar servicios V2X en infraestructuras de red del futuro. Así, como se ha mencionado anteriormente, este trabajo cierra el compendio del doctorado con una prospectiva del panorama de las NGNs, englobando los diferentes aprendizajes y experiencias adquiridas durante el desarrollo de la tesis.

1.4. Conclusiones y Trabajos Futuros

Las NGNs y, especialmente, el 6G, son actores fundamentales en la configuración de la sociedad de los próximos años. La confluencia de los dominios físico y digital formará parte del día a día de millones de personas en todo el mundo. Aunque aún se están definiendo las aplicaciones y servicios del futuro, la telepresencia, la realidad aumentada y virtual, la telemedicina, la conducción autónoma o remota se convertirán en parte esencial de nuestras actividades cotidianas. El 6G debería ser el habilitador para transformar nuestro mundo físico en uno digital, donde se mantendrá una versión muy detallada de nuestra vida. Gracias a este gemelo digital, se podrán realizar diferentes análisis y predicciones para reaccionar en tiempo real ante acontecimientos a punto de suceder. La infraestructura de red soportará las comunicaciones entre ambos dominios, con el apoyo de dispositivos alimentados por IA repartidos a lo largo de todo el continuo de computación y networking, desde los dispositivos finales hasta la nube.

Para hacer posible la consecución de estos hitos, los modelos de IA se convertirán en la piedra angular de la 6G. Los enormes requisitos de red y procesamiento que se necesitarán hacen imposible que los humanos alcancen el nivel de capacidad de gestión y orquestación necesario para gestionar las comunicaciones ubicuas y de gran ancho de banda de los próximos servicios. El proceso de toma de decisiones estará completamente automatizado mediante técnicas de ML, reduciendo la intervención humana a cero, minimizando las interrupciones del servicio a la hora de flexibilizar la red y hacerla resistente a las amenazas de ciberseguridad. De este modo, será posible lograr sistemas basados completamente en ZSM.

Las arquitecturas de red también tienen que sumarse a esta transformación para sentar las bases de la evolución prevista. La flexibilidad y la compartición de recursos serán puntos críticos, ya que se espera una amplia diversificación de servicios con la introducción de múltiples partes interesadas con diferentes necesidades y requisitos. La softwarización de las tareas de red y computación está exigiendo mucha universalidad a los componentes de hardware subyacentes para dar cabida a aplicaciones o servicios con necesidades heterogéneas. Aunque las exigencias a la infraestructura física son altas, la sostenibilidad y la eficiencia energética son de suma importancia para las próximas décadas, tal y como se acordó en el Acuerdo Verde europeo[8]. De este modo, la eficiencia será clave tanto para aprovechar al máximo los dispositivos hardware como para reducir el consumo energético de forma óptima. Mediante la introducción de nuevas tecnologías capaces de realizar tareas tradicionalmente pesadas en dispositivos de recursos limitados, el espectro computacional podría distribuirse aún más, permitiendo así el uso de dispositivos de bajo consumo para procesos que normalmente necesitaban hardware dedicado de alto coste.

En consecuencia, serán necesarios muchos esfuerzos de investigación en todas las tecnologías clave que componen el ecosistema 6G para alcanzar los objetivos mencionados. Estas actividades deben trabajar de forma coordinada para avanzar conjuntamente hacia la realización de infraestructuras de red totalmente basda en ZSM, desde la radio a la nube, y desde los usuarios finales a los servicios, estén donde estén. Con el fin de dar un primer paso en esta dirección, el objetivo principal de esta tesis doctoral ha sido diseñar, implementar y evaluar herramientas novedosas y ligeras de gestión de tráfico basadas en IA para NGNs. En una etapa inicial, se identificó que los algoritmos de ML tienen un papel clave para la integración de la inteligencia en la infraestructura de red (Objetivo 1). Por ello, se realizó una exhaustiva revisión bibliográfica de los modelos adecuados para ser adoptados en estos escenarios. Se estudiaron múltiples trabajos de investigación para identificar los algoritmos más apropiados para las tareas de red individuales que conforman cualquier sistema ZSM [115] (Objetivo 2). Además, se estudiaron las actividades de estandarización relacionadas, ya que impulsan las tendencias futuras en el diseño y desarrollo de NGNs. También se analizaron los esfuerzos destacados en forma de proyectos de colaboración internacional, lo que evidencia el potencial del paradigma ZSM y el interés suscitado tanto en el mundo académico como en la industria. De esta forma, como resultado, se obtuvo una visión general clara de los algoritmos y tecnologías de ML que pueden ser candidatos a ser usados en infraestructuras de red basadas en ZSM.

Para permitir el procesamiento rápido de paquetes en redes complejas es necesario utilizar tecnologías eficientes con gran flexibilidad y notable rendimiento. De este modo, se realizó un estudio de las alternativas existentes para la programabilidad del plano de datos con el fin de seleccionar los candidatos ideales, considerando tecnologías de vanguardia que puedan gestionar grandes cantidades de tráfico de forma eficiente. Las elegidas fueron P4 y eBPF. Para evaluar sus capacidades, se llevó a cabo una extensa evaluación de rendimiento en un entorno 5G multi-operador [116]. Ambas tecnologías se probaron desarrollando funciones de red que proporcionaban inspección profunda de paquetes y aplicaban acciones de calidad de servicio a los flujos de tráfico. Se evaluaron en términos de ancho de banda máximo permitido, tiempo de procesamiento de paquetes y uso de recursos informáticos. Como resultado de estos experimentos, P4 obtuvo un rendimiento óptimo al no mostrar pérdidas de paquetes a altas velocidades de datos y mantener una latencia muy baja en el manejo de paquetes. Por otro lado, eBPF sufrió para gestionar paquetes pequeños con rapidez aunque mantuvo niveles similares de capacidad de procesamiento de paquetes en escenarios con un tamaño medio de paquete más elevado. Estos resultados eran los esperados, ya que P4 está específicamente diseñado para estos fines y opera en equipamiento dedicado y específico, mientras que eBPF es una tecnología más flexible y universal que funciona en hardware genérico. Después de analizar estos experimentos en el contexto de escenarios 5G multi-operador, se llegó a la conclusión de que eBPF es una herramienta adecuada para las redes B5G, ya que puede implementar funciones de red ligeras situadas en cualquier punto de la infraestructura. Además, la inversión necesaria para desplegar componentes basados en eBPF es insignificante si se compara con la inversión necesaria para instalar equipos habilitados para P4. En

[8] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en

esta línea, dada la flexibilidad exhibida, eBPF demostró su capacidad para dar cabida a la integración procesos inteligentes basados en ML dentro de los sistemas ZSM (Objetivo 3).

De esta forma, dotar de inteligencia a las funciones de red basadas en eBPF es un paso necesario para integrarlas en infraestructuras de red totalmente autónomas. Aunque existen varios algoritmos de ML a considerar, las redes neuronales han demostrado ser muy eficientes para escenarios de inspección de paquetes. Esta conclusión se extrae del análisis inicial del estado del arte que se realizó sobre la aplicación de soluciones basadas en ML para tareas de red, que demostró el extendido uso que existe de redes neuronales en la literatura para el análisis de paquetes con excelentes resultados. Por ello, se construyó un modelo de red neuronal para detectar ataques en redes IoT, donde el uso de dispositivos con pocas capacidades está muy extendido y sus recursos computacionales son limitados. Este es el escenario perfecto para evaluar las capacidades de eBPF para integrar algoritmos de ML en hardware genérico e inspeccionar de forma inteligente el tráfico para reaccionar en tiempo real ante ciberataques. Para poder ejecutarse en el entorno eBPF, la implementación de la red neuronal entrenada tuvo que ser ajustada para ser aceptada por el estricto verificador de eBPF antes de su inyección en el Kernel de Linux (Objetivo 4). Se trata de una tarea compleja dadas las severas restricciones impuestas por dicho verificador. Además, a efectos comparativos, se desarrolló otra versión de esta función de red en la que la red neuronal se ejecutaba en el plano de usuario, fuera del entorno de ejecución eficiente de eBPF dentro del espacio del kernel. Tras la evaluación de rendimiento realizada, la solución que implementaba el procesamiento del plano de datos y el algoritmo de ML en el kernel de Linux mejoraba el consumo de CPU y reducía drásticamente el tiempo necesario para ejecutar la red neuronal. En consecuencia, se demostraron las capacidades de eBPF para operar de forma flexible en el extremo lejano de la red, sobre dispositivos restringidos, aportando una toma de decisiones inteligente en puntos de la red nunca antes vistos [117] (Objetivo 5).

Finalmente, como última aportación de la tesis, se puso en práctica toda la experiencia y conocimientos adquiridos en infraestructuras de red B5G y arquitecturas basadas en ZSM para vislumbrar cómo los servicios del futuro darán forma al diseño de infraestructuras 6G [118]. Se tomó como base el vertical vehicular, ya que exige unos estrictos requisitos de red y de computación para soportar la próxima oleada de servicios vehiculares. De esta forma, las aplicaciones del futuro que se espera que demanden unas capacidades de red actualmente inalcanzables, e incluso no concebidas todavía, se distinguieron y se situaron en el estado del arte (Objetivo 7). A partir de este estudio, se identificó un conjunto de KPIs medibles, y se presentaron como una lista de indicadores que pueden utilizarse para determinar el rendimiento de la infraestructura de red de forma objetiva (Objetivo 6). Los KPIs se clasificaron en función del segmento de red donde son relevantes, de forma que se pueda evaluar de forma concreta si la infraestructura desplegada es capaz de satisfacer las demandas de los servicios futuros. Además, el estudio realizado durante todas las etapas de investigación de esta tesis doctoral con respecto a las arquitecturas ZSM y B5G se puso en práctica mediante el diseño de una plataforma conceptual 6G para gestionar aplicaciones en red. Esta plataforma se basa en los principios de ZSM con un punto de vista experimental, donde todo el ciclo de vida de los servicios desplegados debe ser monitorizado y evaluado automáticamente para valorar su rendimiento. Esto es algo fundamental para los escenarios vehiculares en redes B5G, ya que una parte esencial de ellos son los servicios críticos. Estos servicios deben de gestionarse de manera meticulosa, algo que solo es posible con el uso de este tipo de plataformas.

Es importante destacar que los resultados obtenidos durante el desarrollo de esta tesis doctoral, incluyendo la implementación de la solución basada en ML y eBPF para el procesamiento eficiente de paquetes en entornos B5G, están siendo explotados y reusados en diversos proyectos de investigación europeos. Por ejemplo, en RIGOROUS[9] (código de proyecto: 101095933), donde el objetivo es introducir mecanismos de seguridad basados en IA para reaccionar dinámicamente a amenazas en la capa de orquestación y a las funciones de red. De la misma forma, en el proyecto HORSE[10] (código de proyecto: 101096342), también se utilizan este tipo de servicios inteligentes para habilitar la predicción de ciberataques, la programabilidad del plano de datos, y la orquestación y gestión autónomas.

[9] https://cordis.europa.eu/project/id/101095933
[10] https://cordis.europa.eu/project/id/101096342

Aunque los resultados presentados en esta investigación pueden servir de base para el desarrollo de herramientas ligeras de procesado de tráfico basadas en IA para infraestructuras de red B5G, aún quedan múltiples líneas de investigación por explotar para seguir avanzando en el estado del arte. En primer lugar, el uso de eBPF para implementar funciones de red inteligentes se encuentra aún en sus fases iniciales, y todavía no se ha alcanzado el máximo potencial de esta sinergia. Un conocimiento más profundo de la tecnología y mayores esfuerzos de desarrollo crearán infraestructuras de comunicación más eficientes y potentes, capaces de hacer frente a requisitos de aplicaciones y servicios más exigentes. Además, el principal inconveniente del entorno eBPF son las limitaciones impuestas por el verificador para asegurar la correcta ejecución y terminación de los programas en el contexto privilegiado. Así, desarrollando algoritmos más optimizados de forma automatizada, será posible exprimir su rendimiento e implantación hasta el siguiente nivel. Asimismo, seguir esta línea de investigación conducirá a la posibilidad de integrar más modelos de ML, permitiendo la creación de una gama diversa de funciones de red inteligentes que puedan ejecutarse de forma flexible en cualquier punto de la infraestructura. De este modo, será posible dotar de inteligencia y autonomía a todo el continuo de red y computación y aumentar el rendimiento global de las infraestructuras de red 6G. En esta línea, la nueva ola de modelos LLM tendrá una importancia destacada en este proceso, ya que permiten adquirir rápidamente una gran cantidad de conocimientos, por ejemplo, especificaciones de protocolos, mejores prácticas, estándares, configuraciones de dispositivos de red existentes, etc. Esto va más allá de las capacidades humanas y de los actuales sistemas de gestión de redes. La integración de los LLM en las infraestructuras de red permitirá unificar la inteligencia de red a través de del procesamiento del lenguaje natural para automatizar la gestión y orquestación de la red, cerrando la brecha entre los administradores de red y la configuración de la infraestructura de red subyacente.

Por último, otra línea de investigación de interés sería la integración de estas funciones de red en el marco 6G propuesto para cerrar el bucle y obtener una arquitectura de red ZSM plenamente funcional. Esto permitirá la construcción de servicios más complejos a través de diferentes dominios utilizando patrones de interoperabilidad. Además, las funciones de red podrían integrarse en los procedimientos de testeo y validación para evaluar su funcionamiento tras el despliegue y decidir si se cumplen los requisitos de red exigidos. Esta plataforma también supervisa la calidad de servicio proporcionada en la infraestructura 6G y puede reaccionar ante disminuciones del rendimiento. Como respuesta, puede activar la reconfiguración de los recursos de cómputo y de red, o la aplicación de procedimientos de migración en tiempo real a una ubicación más adecuada a lo largo del continuo de computación y de red.

CHAPTER 2

Abstract

2.1. Motivation

The advent of NGNs will bring challenging demands for the communication infrastructures supporting them. They will require more flexibility and autonomous orchestration, thanks to the integration of AI at different levels, in order to achieve higher speeds and lower latencies, among other KPI of interest [1,2]. These networks will become key enablers for futuristic and disruptive services across multiple vertical segments, e.g., Industry 4.0, autonomous and cooperative driving, holographic/telepresence services, etc. [3]. To handle the upcoming explosion of new applications, several network management solutions are surfacing to cope with the high complexity of the dissagregated architecture envisioned for B5G systems. This is because the services and applications of the future will demand unprecedent performance levels from the underlying communication infrastructures and new network management workflows unattainable by humans without the help of AI-based mechanisms. In addition, to fulfill all this requirements and address the user needs, the transition to NGNs implies a tremendous investment in new physical infrastructure and, therefore, in updated hardware and software solutions.

During the last two decades, many frameworks providing network management and orchestration mechanisms have been proposed and implemented following different approaches. However, although there have been several successful solutions used in production environments, there is still a necessity of improving them to satisfactorily support the upcoming B5G applications and services, as mentioned before. Network programmability and flexibility arise as two of the fundamental pillars of NGNs, hence they should drive the development of future network functions [4]. In this line, SDN and NFV have emerged during the last years as promising technologies to address the needs of future network infrastructures, although their maximum potential is still to be achieved. While these technologies have provided new dimensions in network flexibility and programmability, the softwarization of the network demands more innovative solutions. The aim is to address an autonomous and high performing management and orchestration of network resources in order to meet end-user's demands in terms of QoE. Furthermore, the landscape is becoming more complex with the ever increasing appearance of multi-stakeholder networks virtually sharing the same physical infrastructure as proposed in the envisioned 6G paradigm [4]. Under this vision, in the new era of communications, there will be billions of humans, vehicles, and tiny things interconnected and generating huge amounts of data through heterogeneous network infrastructures owned and managed by different actors: Telcos, infrastructure operators, service providers, etc.

NGNs should lead the way in the upcoming years to achieve distributed, autonomous, sustainable,

flexible and trustworthy computation. From the governance angle, AI is the cornerstone to maintain cost-effective maintenance of the forthcoming complex services operating in the networks of the future. The softwarization of the network have just started with the proliferation of a myriad of virtualization technologies, and their interoperability will pose a formidable challenge for network operation. Thus, AI-powered management and orchestration functions will be the key to automate the decision-making process. In this way, ML is being adopted as a crucial technology to provide the required intelligence to these systems. ML-based techniques will provide the network infrastructure with autonomous governance abilities to make it self-adaptable considering its own conditions and user/service needs in real-time. The evolution of AI-based network functions will enable predictive orchestration to optimize even more traffic management, resource placement and service configuration based on expected needs, anticipating the demands of the users. This is what will permit the realization of the ZSM, where the human intervention will be reduced to zero as the network will start to operate autonomously; hence, its efficiency will be optimized to levels that have never been seen before.

It is also envisioned that the traditional networking and computing architectures that are currently evolving to the MEC paradigm will converge to an unique computing continuum covering from the cloud down to the end users or devices (cloud–core–edge–end-device continuum). As a result, ubiquitous computing will democratize the use of resources among all the actors making use of the network, hence the response capacity will drastically increase. Besides, the volume of the typical client-server traffic flows will be reduced, decreasing the latency of the network. Thanks to the trend to virtualize the services as chain of network functions, it will be easier than ever to distribute the networking and computing tasks all over the continuum. The significance of this paradigm is even greater if we consider the NGNs' objective of achieving zero latency, which is only possible with the introduction of local or close-to-the-user processing.

This whole revolution brings along a notable investment in equipment, thus reducing its associated expenses, i.e., CAPEXs, is also of utmost importance. Besides, improving the energy efficiency of the network and reduce the ecological footprint is also crucial for our future. One of the ways to address these issues is by exploiting and reuse inexpensive hardware and introducing lightweight virtualization and fine-grained resource management, while maintaining high network and computing performance levels. This implies the adoption of new virtualization technologies and service composition models to leverage the new network management and orchestration paradigms. Moreover, it opens the door for resource-constrained and non-specialized devices to participate in the network infrastructure operations. The aforementioned unique computing continuum will benefit this approach, as both control plane functionalities and data plane programmability will be seamlessly managed in an integrated basis along this end-to-end continuum. Furthermore, the expected traffic speeds that will be achieved in NGNs will demand ultra-high efficient packet processing capabilities from the commodity hardware composing the far edge in order to bring over the intelligence as close as possible to the users and reduce the latency to the lowest. To this end, data plane reprogrammability techniques will have to be part of this transformation to become the foundations of the network architecture on top.

One of the vertical markets where this B5G paradigm evolution will impact the most is the vehicular one. During the last years, V2X communications harvested a lot of interest, from both industry and academia. The pictured improvement of road safety, carbon footprint decrease, introduction of holographic display systems, immersive infotainment, and enhanced user experience demands a disruption in the way vehicular networks are conceived. V2X envisage a lot of actors involved in a dynamic environment, at the time of considering a wide range of connectivity strategies to cover all the scenarios, e.g., V2V, V2P and V2I [5]. To effectively provide a proper QoE to all the involved stakeholders, V2X will push to the limits the capacities of existing network architectures. That is the reason why it needs NGNs in conjunction with ZSM as a bedrock. Besides, due to the extreme heterogeneity present in the automotive ecosystem, in terms of communication scenarios, rigorous requirements and limited computing capabilities, new solutions are needed to enable autonomous and flexible decision making while sustaining a high performance.

While it is true that the research community has deeply investigated these topics and provided significant advances in the challenges identified during the last years [6], there are still some dots to be

connected and multiple gaps in the state of the art. These were the seeds that originated this PhD journey. In the beginning point of the thesis, ML techniques were starting to be used to enrich network architectures with intelligence and autonomous management, although they were in their infancy stage and, therefore, still facing several challenges. In this way, their application was studied as a limited solution to handle only certain segments of the network by using very specific ML models. However, this restricted vision needed to be expanded to explore the holistic and automatic management of complex network infrastructures. As proposed by the ZSM paradigm, this can be only achieved by analyzing the problem from two different, although coupled, perspectives: The data plane management and the control plane administration. The former requires fast and efficient technologies to process the massive amount of traffic that will cross the network infrastructure in NGNs. In this way, there were a lack of studies exploring and validating these traffic-handling technologies in real environments to select the best solutions for each scenario, considering the unique computing and network continuum encompassing from the end-devices to the cloud. In a similar manner, the literature barely covered the integration between data plane reprogrammability techniques and intelligent traffic handling models, evidencing the necessity to automatically manage both control plane functionalities and data plane processing in a seamless way. Besides, the validation of these solutions was usually made considering the availability of high computing capabilities, leaving behind a plethora of devices located in the edge and the far edge of the network. Moving on to the design of the infrastructures of the future, generalist approaches were the trend to cover the multiple dimensions related to the evolution of the communication networks. Additional concretion was needed to extract tangible and objective metrics to be fulfilled by B5G infrastructures. Finally, there was also a necessity of management architecture designs to handle the upcoming applications and services in an efficient way, while making the most of the underlying network infrastructure and its resources.

The present PhD thesis describes the research results of the design, implementation, and validation of network reprogrammability solutions fueled by ML, as enablers for the fulfillment of the ZSM paradigm in NGNs infrastructures. The proposals in this thesis permit the integration of AI models within data plane management schemes in commodity hardware, aiming at accommodating the necessities of the upcoming computing and networking continuum. The data plane reprogrammability is provided thanks to eBPF [1]. This technology permits the execution of user-developed programs within the Linux kernel, enabling faster packet processing, hence, introducing minimal disturbance to traffic flows. Furthermore, it also allows the integration of complex AI algorithms traditionally executed in dedicated servers into commodity hardware. While the range of ML algorithms is wide and several could have been considered, NNs have been adopted, as they provide great flexibility and high accuracy to analyze traffic and detect anomalies [7]. With these powerful tools, this thesis explores two timely and different scenarios, namely, (i) high-volume traffic processing in a 5G network for QoS purposes, and (ii) integration of advanced traffic analysis in constrained IoT devices with focus on cybersecurity. To wrap everything up, all the acquired knowledge is put into practice to envision how future applications and services will direct the design and realization of NGNs architectures. This exercise is canalized through the vehicular vertical, as it is setting the upcoming high demanding required to be tackled by the underlying network infrastructures. As a result, critical KPIs are identified to provide objective and measurable metrics to serve as baseline performance requirements to be addressed by B5G infrastructures. Finally, a holistic platform is presented aiming to establish the building blocks that will support the future needs of NGNs.

This PhD was supported by the Fundación Séneca—Agencia de Ciencia y Tecnología de la Región de Murcia (Spain)[2]—under the FPI Grant *21429/FPI/20*, and co-funded by Odin Solutions S.L.[3], Región de Murcia (Spain); and by the Fulbright Commission in Spain[4] under the Fulbright grant *00003/FLB/21*.

[1] https://ebpf.io/
[2] https://fseneca.es/
[3] https://odins.es/
[4] https://fulbright.es/

2.2. Goals and Methodology

To advance the state of the art and address the challenges discussed previously, the main objective of this thesis is to explore the current needs of NGNs to adopt fully functional ZSM capabilities. It is intended to provide the integration of different technologies that permit the fast management of the data plane, e.g., efficient packet processing, directed by AI-driven network functions enabling the autonomous management of the network in real-time. The followed methodology focused on addressing the convergence of data plane programmability technologies and AI schemes. Therefore, in first place, this thesis focuses on a meticulous study to identify key ML algorithms well positioned to be undertaken by ZSM systems. Next, it moves the lights to the search, selection, and performance evaluation of technologies that allow fast packet processing with near to zero impact in the network performance. Then, both parts are merged to achieve data plane reprogrammability directed by intelligent models that enable the creation of tools well-suited for ZSM-based NGNs infrastructures. Their feasibility is demonstrated by validating their execution in commodity hardware, showcasing the integration of traditionally resource-demanding computing processes in lightweight equipment. Finally, the acquired knowledge and expertise is exploited to envision key aspects in the design of 6G infrastructures in the context of the vehicular vertical, which poses severe network challenges to satisfy its stringent requirements for the development of V2X services. To achieve the main research goal of this thesis, the following objectives were identified:

- **Objective 1**: Analyze and study the state of the art in the application of ML-based techniques to network management and orchestration tasks.
- **Objective 2**: Identify adequate and efficient ML algorithms well positioned to be adopted by ZSM systems in complex network infrastructures.
- **Objective 3**: Select network programmability technologies suitable to address NGNs challenges.
- **Objective 4**: Implement and integrate ML-powered traffic processing techniques in B5G infrastructures to provide autonomous and fast packet handling with no performance degradation.
- **Objective 5**: Evaluate the developed solutions in commodity hardware to enable their deployment in the whole computing and network continuum, from the end-devices to the cloud.
- **Objective 6**: Capitalize the acquired expertise to envision the key points that will direct the design of 6G infrastructures.
- **Objective 7**: Explore how the upcoming and high-demanding services and applications will defy the 6G networks in the context of a stringent and challenging vertical such as the vehicular one.

In order to reach those objectives, the work performed during this PhD was divided into different research lines corresponding to each one of the goals. After completing all of them, they converged as a whole to compose this doctoral thesis. In first place, there was a search and study phase, in which the literature was reviewed to deeply understand the state of the art of the ZSM concept, ML applied to networking, standardization activities, and ongoing research projects for NGNs. Once this was done, the tools and technologies most prominently used in these areas were explored and examined, with the aim of selecting the most proper ones to achieve intelligent network reprogrammability in ZSM scenarios. eBPF was the selected packet-handling technology after experimentally comparing it to P4. An extensive performance evaluation was conducted to obtain insights of their capacity to process traffic at high speed, in terms of maximum permitted bandwidth, latency and CPU usage; evidencing that eBPF is a flexible, simple and cost-effective solution that may be deployed in multiple scenarios without presenting prohibitive hardware requirements. By using eBPF as a traffic processing technology, the next stage was to select the proper ML algorithms to analyze the traffic and automatically react to changes, disruptions, or threats in the network. The initial literature review revealed that NNs

are one of the fundamental pillars to enable smart decision-making in ZSM environments. In this way, a development was made to combine eBPF packet processing with MLPs to obtain ML-enabled networking capabilities. As a result, a packet handling pipeline based on ML was integrated within the Linux kernel, thus accelerating the efficient execution of AI algorithms. To validate the developed implementation, a real-world use case was explored, in which the solution examined real-time traffic while using a NN to detect attacks in an IoT network. Finally, with all the knowledge acquired during the development of the aforementioned research lines, a final exercise was made to envision the key networking and computing aspects that will drive the design of future 6G infrastructures. The stringent and challenging automotive vertical was studied from the point of view of its future applications and services to explore how they will shape the 6G network's definition and deployments.

2.3. Results

The completion of each one of the proposed objectives of this PhD thesis produced several scientific publications. Four of these publications have been selected to compose this PhD thesis compendium. It is worthy to note that these articles have been published in international journals ranked in the first quartile of the Clarivate Analytics' JCR [5]. The key results obtained during this PhD thesis, as outcomes of the performed research, are presented in Table 2.1, in which their associated objectives and publications are also indicated.

Table 2.1: Main thesis results

Result	Objectives	Publications
R1. Analysis of the state of the art ML algorithms and techniques well suited to be adopted by ZSM systems.	1, 2, 4	[115], [117]
R2. Analysis of cutting-edge technologies used for fast packet processing in B5G networks.	3, 4, 5	[116], [117]
R3. Design, implementation and validation of a QoS-oriented solution based on eBPF and P4 that operates in multi-tenant 5G networks.	2, 3	[116]
R4. Extensive performance evaluation of eBPF and P4 in terms of packet processing performance and programmability features.	3, 5	[116], [117]
R5. Design and implementation of a ML-based pipeline for efficient packet-processing within the Linux kernel.	1, 2, 4, 5	[115], [117]
R6. Integration and validation in commodity hardware of an eBPF-based solution that enables ML intelligence to dynamically react to attacks in an IoT environment.	4, 5	[115], [116], [117]
R7. Comprehensive identification of concrete and objective KPIs that can measure the performance of 6G infrastructures.	6, 7	[118]
R8. Design of a ZSM-orchestrated 6G platform capable to manage the whole lifecycle of multi-domain services of the future.	1, 6, 7	[118]

Besides the four journal papers composing this thesis, other related works have been also published: Two book chapters, ten conference articles, and four additional JCR-indexed articles. Note that this thesis has been presented as a compendium of four research articles which represent the core of this research work. Thereby, the main contribution of the conducted research is contained within these four articles and the rest of publications extend the contributions of this thesis to the state of the art.

In the following, the reader can find a detailed summary of the work carried out in each one of the publications that compose this PhD thesis, linking them more explicitly with the obtained results. Besides, the complete documents of these publications are included in Chapter 4.

[5] https://jcr.clarivate.com/jcr

2.3.1. Machine learning-based zero-touch network and service management: a survey

The first work composing the compendium [115] explores ML techniques well-suited to be empower the ZSM paradigm. Through a deep review of the ML algorithm taxonomy, multiple proposals found in the literature to manage and orchestrate NGNs infrastructures are recapitulated and analyzed. As a result of this study, the different network control functions enabling the automatic management of the network are identified at the time of providing understanding of the most adequate ML techniques for implementing each component (**R1**). ML algorithms are usually categorized in four different families: SL, UL, SSL, and RL. The selection of the most suitable algorithm is multi factorial as it depends on the characteristics of the underlying communication infrastructure, the network traffic requirements, the expected behaviour of the users, and the type of services and applications operating over the network. The discussion of the proposals applying AI to the management and orchestration of the network is driven according to the roles of the studied network functions within the network architecture: (i) Flow inspection, (ii) multi-domain management, (iii) RAN management, and (iv) network resource management. In the conducted literature review, it was evidenced that SL techniques are the most used in flow inspection scenarios, as operators usually has their own policies to classify the traffic travelling over their infrastructure, and the use of supervised algorithms can leverage this information and achieve high efficiency. In multi-domain scenarios, hierarchical approaches perform better than centralized solutions due to the partial view of the network that the different components present. Typical ML algorithms find it difficult to efficiently operate in these complex and changing environments as real-time training is very costly because of the amount of samples required. As a consequence, RL solutions are the typical choice to retrieve information from the multiple domains and select the most adequate algorithm to predict the expected behaviour. Maintaining a real-time operation of ML algorithms in the RAN management implies the use of a tremendous amount of computational resources, which makes it very costly due to the dynamic characteristics of the RAN. In this way, although the existing solutions for this segment of the network are based on RL to quickly adapt to changes in the radio medium, the majority of them make an offline training that is later deployed on the production environment. Resource management is a broad field and it covers different aspects in the control of the network, thus there is no an unified way to approach its administration. In the literature, both SL and RL solutions are predominant. The former leverage its capacity to deal with high network complexity and heterogeneity, while the latter provides flexible and real-time reaction to changes in the traffic and user behaviour. Besides this analysis, this work also explores the standardization activities and research projects working towards the definition and evolution of ZSM. The interest that ZSM is attracting in both academia and industry is evidenced by the great number of initiatives providing conceptual solutions and real implementations, which is also deeply discussed in this paper. Finally, the challenges and future lines of the automated management and orchestration of NGNs are discussed. In this way, the work performed in this journal article provides a deep understanding on the current status and future trends in ML-based ZSM and lays the foundations for the next published works forming the compendium.

2.3.2. Fast traffic processing in multi-tenant 5G environments: A comparative performance evaluation of P4 and eBPF technologies

The second work of the compendium [116] introduces a review and a performance evaluation on two of the most promising technologies to handle high bandwidth traffic while ensuring low packet processing latency in NGNs: P4 and eBPF (**R2**). P4 is a programming language targeted at PISA architectures, which is the reference data plane model for programmable network devices. The basis of this architecture is a pipeline where, in first place, the packets bits are aligned with the protocols; then, multiple actions can modify the packet if it matches a series of rules; and, lastly, the bit stream is converted again into packets before send it back to the network. In turn, eBPF is a technology that enables running programs in a sandboxed and safe environment within the Linux kernel. It

permits to load new software inside the kernel on-the-fly and without modifying its code, extending the capabilities of the kernel in a secure way. When combined with the Linux's XDP, it permits the efficient implementation of powerful, while lightweight and portable, network functions. The operating system is the one guaranteeing the safety and execution of the programs with the help of a verification engine and a just-in-time compiler. Both technologies are studied this work, evidencing to be capable alternatives to enable fast packet processing and programmability in NGN infrastructures. In the conducted experiments over a real 5G infrastructure, P4 showcases very high performance due to its hardware-based traffic processing, having a expensive deployment cost as its main drawback. On the other hand, eBPF was still in its inception stages when used in this paper for data-plane reprogrammability, but it shows bright capabilities as a simple, portable and efficient network functions enabler. Despite the promising future of these technologies, prior to this article there were no studies in the literature evaluating and comparing their performance in NGNs. Thus, the purpose of this work is to serve as a performance indicator baseline of P4 and eBPF for the design, development and deployment of network functions in NGN infrastructures with ZSM capabilities. In first place, a discussion about the advantages and drawbacks of both of them is done, together with a literature review addressing the most prominent works in the field. By doing so, it is identified scenarios where one or another is more adequate. Next, a QoS-oriented 5G packet processor is presented using both technologies with the aim of providing a network function able of performing early packet processing in a MEC node (**R3**). Their performance is evaluated from different perspectives in the aforementioned 5G testbed so different-nature results are shown. P4 obtains the best figures in terms of packet processing rate, which is not a surprise; as a hardware-based solution it is expected to outperform any other one based on software artifacts. However, the tuple formed by eBPF and XDP obtains similar results in high-bandwidth scenarios where the mean length of the packets is high. Besides, to evaluate the maximum capacity of these technologies, the experiments conducted use extreme traffic volumes. But, with moderate and more typical traffic loads, the eBPF/XDP solution matches the performance of the one implemented using P4. In the light of these results, there are some scenarios in which the network requirements are not that stringent like in the core of the network, such as MEC use cases or IoT scenarios. In these environments, lightweight and flexible eBPF/XDP network functions can be easily placed on demand in different points of the network. Also, they can be handy in ZSM systems, where the network conditions can change in any point of the infrastructure and these small network functions can be easily deployed with any desired behaviour in a matter of seconds. Besides, when considering the CAPEX of NGN, specialized networking equipment is expensive to deploy in different points of the infrastructure. Nevertheless, deploying commodity hardware capable of hosting eBPF/XDP-based network functions can significantly reduce the cost of the infrastructure, while maintaining fast traffic processing capabilities in a flexible and efficient manner. After completing this study and the performance evaluation of both technologies (**R4**), eBPF erected as a promising technology to enable the use of complex network functions, as those enabled by ML algorithms, in ZSM scenarios. In this way, eBPF was the selected technology to continue the research journey followed in this thesis. The next step was to converge the knowledge and background acquired with the two first works of this thesis. Concretely, the focus was on developing efficient network functions leveraging the synergies between eBPF and ML to perform automated networking tasks at any point of the infrastructure in an effortless way, which is the challenge addressed in the next article of this compendium.

2.3.3. Machine learning-powered traffic processing in commodity hardware with eBPF

The third article of the compendium [117] presents the development and evaluation of a network function that takes advantage of eBPF and ML to provide intelligent processing within the Linux kernel in commodity hardware (**R6**). The implementation enables the execution of heavyweight computational tasks, such as NNs, in a simple and flexible way. As it is discussed in the second article of the compendium [116], eBPF posed as an enabler to develop efficient network management and

monitoring at any point of the infrastructure (**R4**), which is highly valuable to develop security network functions. This is because the traffic inspection can be done by any Linux-enabled device at high line rates with minimal impact in the latency. Therefore, unauthorized accesses and attacks can be detected in real-time. The typical approach found in the literature combines the use of ML algorithms running in user space that are fed with data collected by eBPF probes. Nevertheless, this decoupling can result in a performance degradation, as the user space is less efficient to perform computational tasks when compared with an in-kernel implementation. In this way, the developed network function combines fast packet inspection and ML-based decision making within the Linux kernel, saving computational resources from the device and reducing the processing latencies. Following this approach, it is possible to increment the performance of intelligent networking tasks in a flexible and portable way, which is of utmost importance in NGNs. The implemented solution is validated in an IoT use case, where the use of devices with constrained computational resources is the norm. The limited capabilities of these devices in terms of memory, processing power, and energy consumption pose a challenge for embedding intelligent networking or cybersecurity functions. In this environment, every node is a possible entry vector to the whole network, thus strengthening their defense capabalities against cyberattacks is critical for the robustness of the infrastructure. Thanks to the eBPF efficiency to run code at Linux kernel level, sophisticated security solutions that would be heavyweight to run in a typical context, can be fully integrated within the operating system; hence, enabling a new range of security network functions to be executed in commodity hardware. In addition, the inclusion of intelligent algorithms at this level also permits the automatic detection and reaction to security threats in IoT environments in a flexible and lightweight manner. Concretely, in the considered use case, the developed eBPF network function is validated in a 6LoWPAN network using RPL. 6LoWPAN is a standard defined by the IETF to integrate IPv6 in low-power wireless networks, such as those defined by the IEEE 802.15.4 standard [8]. RPL is a routing protocol for this kind of networks based on distance vectors that operates over 6LoWPAN. It is one of the most used IoT routing protocols due to its facility to create new routes, share routing information, and adapt to topology changes. In this way, the evaluated scenario is formed by 6LoWPAN nodes threatened by the *Hello Flood* attack on the RPL protocol. In this case, the attack consists on several malicious devices sending routing information requests to the rest of the nodes, forcing them to respond and wasting their resources. The first step to evaluate the performance of the developed network function is to generate a dataset to train the ML models. To do so, an IoT network is deployed using the Cooja simulator[6], a tool that permits the emulation of virtual devices implementing 6LoWPAN and RPL. In this environment the attack is simulated and three datasets are obtained. Next, although multiple ML algorithms could have fit in this use case, the use of MLP was the final choice, due to its good performance and efficiency to inspect traffic and detect anomalies [7]. However, MLPs are complex models and cannot be directly implemented in eBPF because of the constrains imposed by the eBPF verifier. Therefore, the starting point was to use the Python's Scikit-Learn MLP implementation[7]. The model was trained with a 80/20 split and, after multiple tests, a MLP model conformed by two hidden layers (composed of three and two perceptrons, respectively) was obtained (**R5**). The model attained an accuracy between 0.9 and 1 in all the considered scenarios, although the focus of the work is not achieving great accuracy in the attack detection but running the MLP model inside the Linux kernel, something not achieved until this point in the literature. With the model ready, the TinyML *emlearn* library [9] was used to convert the Python code into C code, which was the basis of the eBPF program. This code was then modified and customized to be validated by the eBPF verifier and embedded in the Linux kernel. Another version of the network function was also developed to decouple its funtionality, i.e., locating the packet parsing at kernel level, and the execution of the ML model for its deep inspection at user space, for comparison purposes. The performance evaluation conducted on a real IoT device shows a clear improvement of the full in-kernel solution, improving in a 6% the CPU usage and reducing the time needed to run the MLP algorithm in a 97%. This outcome demonstrates the advantages resulting from the synergies between eBPF and ML in the development of complex and intelligent

[6] https://github.com/contiki-os/contiki/wiki/An-Introduction-to-Cooja
[7] https://github.com/scikit-learn/scikit-learn

network functions able to run in commodity hardware at any point of the network and computing continuum. These findings also pave the way for the exploration of the next-to-come paradigm in communication infrastructures, namely, 6G, which is the main focus of the following paper in this PhD thesis compendium.

2.3.4. The role of vehicular applications in the design of future 6G infrastructures

Finally, the fourth and last article of the compendium [118] aims at applying all the knowledge acquired during the PhD about NGNs to the next breed of cellular communications infrastructures and focused on the automotive vertical. This work addresses the 6G-enabled V2X from a service-centric approach, providing a comprehensive overview of the forthcoming vehicular ecosystem and its integration with 6G. The design of the 5G architecture was directed by a network-centric approach, which initially lacked concrete requirements of the applications covered by the three service families, i.e., uRLLC, mMTC, and eMBB. In the upcoming definitions of 6G, it becomes of paramount importance to consider a service-centric view from the start to align the design of future infrastructures with the demands of applications operating over them. Although 6G aims at covering a wide spectrum of vertical sectors, it is necessary to specifically study all of them, fleeing from generic approaches. In this way, vehicular applications are a good example of an important vertical that will demand specific and stringent requirements to future 6G infrastructures. The characteristic use cases in this vertical demand low latency and high bandwidth in critical scenarios, as well as robustness to disconnections due to their inherent mobility [10]. In this line, it is crucial to identify the concrete requirements of these services to define how the NGN infrastructures will be built. From the conducted study reviewing both the evolving 5G applications towards 6G, and the envisioned disruptive services that will arrive in the future, the performance requirements that they will pose to the upcoming 6G infrastructure are identified as follows: (i) Deterministic and anticipated end-to-end QoS, (ii) different access technologies, (iii) AI-powered network orchestration and reprogrammability, (iv) AI-based application life-cycle management, (v) multi-domain secure and reliable application onboarding environments, (vi) multi-dimension perception and accurate positioning at high-speeds, (vii) data caching and data freshness, (viii) data privacy and security, and (ix) energy efficiency. All of them serve as a basis for the specific and measurable proposed KPIs that can capture the performance of the network infrastructure (**R7**). Besides, and based on the analysis, the design of a ZSM-based 6G platform is also presented to handle the whole lifecycle of multi-domain V2X services and applications (**R8**). The platform is based on an experimental perspective, where the design and development of future network services should automatically go through a validation stage in testing infrastructures before its deployment in production scenarios. The proposed platform enables the validation of the applications in compliance with certain requirements that come from specific KPIs. Furthermore, the design complies with AI-powered ZSM with the aim of handling V2X services in NGNs infrastructures. Thus, as mentioned previously, this paper closes the PhD compendium with a foresight of the NGN landscape, encompassing the different learnings and experiences acquired during the development of the thesis.

2.4. Conclusions and Future Work

NGNs and, specially, 6G, are fundamental actors on the shaping of the society for the years to come. The confluence of the physical and the digital domains will be part of the workaday of millions of people all over the world. Although the applications and services of the future are still being defined, telepresence, augmented and virtual reality, telemedicine, autonomous or remote driving will become an essential part of our daily activities. 6G should be the enabler for transforming our physical world into a digital one, where a highly detailed version of our life will be maintained. Thanks to this digital twin, different analysis and predictions could be done to react in real-time to events

about to happen. The network infrastructure will support the communications between both domains, supported by AI-powered devices embedded through the whole network and computing continuum, from the end-devices to the cloud.

To enable the achievement of these milestones, AI models will become the cornerstone of 6G. The enormous network and processing requirements that will be needed make it impossible for humans to catch up with the level of management and orchestration capabilities needed to handle ubiquitous and high-bandwidth communications of the upcoming services. The decision-making process will be completely automated using ML techniques, reducing the human intervention to zero, minimizing the service disruptions at the time of making the network flexible and resilient to cybersecurity threats. In this way, it will be possible to accomplish pure ZSM-based systems.

Network architectures also have to join this transformation to provide the foundations of the expected evolution. In the network, flexibility and resource sharing will be critical points, as it is expected a wide diversification of services with the introduction of multiple stakeholders with different needs and requirements. The softwarization of the networking and computing tasks is demanding a lot of universality from the underlying hardware components to accommodate applications or services with heterogeneous necessities. Although the demands to the physical infrastructure are high, sustainability and energy efficiency are of utmost importance for the decades to come, as agreed in the European Green Deal[8]. In this way, efficiency will be key both to make the most of the hardware devices and to reduce the energy consumption in an optimal way. By introducing new technologies capable of performing traditionally heavy computational tasks in resource-limited devices, the computing spectre could be distributed even more, hence, enabling the use of low-consumption devices for processes that usually needed costly dedicated hardware.

In consequence, lots of research efforts will be required in all the key technologies composing the 6G ecosystem to reach the aforementioned goals. These activities should work in a coordinate way to advance together towards the realization of fully ZSM network infrastructures, from the RAN to the cloud, and from the end-users to the services, wherever they are located. In order to put an initial step towards this direction, the main goal of this PhD thesis research period has been to design, implement, and evaluate novel and lightweight AI-based traffic management tools for NGNs. In an initial stage, ML algorithms were identified to have a key role for the integration of intelligence in the network infrastructure (**Objective 1**). For that reason, it was conducted a comprehensive literature review of ML models suitable to be adopted in these scenarios. Multiple research works were surveyed to identify the more appropriate models for the individual networking tasks conforming any ZSM system [115] (**Objective 2**). Besides, the related standardization activities were also studied, as they drive future trends on the design and development of NGN. Prominent efforts in the form of international collaborative projects were analyzed as well, evidencing the potential of the ZSM paradigm and the interest raised in both academia and industry. Thus, a clear overview of the ML algorithms and technologies well positioned to be embraced by ZSM-based infrastructures was obtained as a result.

To enable fast packet processing in complex networks it is necessary to use efficient technologies with great flexibility and notable performance. In this way, a study of the existing alternatives for data plane programmability was performed to select the ideal candidates, considering state of the art technologies able to efficiently handle high volumes of traffic. The choices were P4 and eBPF. To assess their capabilities, an extensive performance evaluation in a multi-tenant 5G environment was carried out [116]. Both technologies were tested by developing network functions that provided deep packet inspection and applied QoS actions to traffic flows. They were evaluated in terms of maximum permitted bandwidth, packet processing time, and computing resources usage. As an outcome of these experiments, P4 obtained optimal results by not showing packet losses at high data rates and maintaining a very low packet handling latency. On the other hand, eBPF suffered to manage small packets fastly although it maintained similar levels of packet processing capacity in scenarios with a higher mean packet size. This behavior was an expected result, as P4 is specifically designed for these

[8] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en

purposes and works on networking equipment, while eBPF is a more flexible and universal technology working on commodity hardware. After the performed analysis in the context of multi-tenant 5G scenarios, it was concluded that eBPF is an appropriate tool for B5G networks as it can implement lightweight network functions located at any point of the infrastructure. Besides, the CAPEX required to deploy eBPF-based components is negligible when compared with the investment needed to install P4-enabled equipment. In this line, given the flexibility exhibited, eBPF demonstrated its capacity to accommodate the integration of ML-based decision making within ZSM systems (**Objective 3**).

In this line, supplying eBPF-based network functions with intelligence is a necessary step to integrate them in fully autonomous network infrastructures. Even though there are several ML algorithms to be considered, NNs, and concretely MLP, have proven to be computationally efficient in packet inspection scenarios. This decision is made after the the initial analysis of the state of the art in the application of ML-powered solutions for networking tasks, which demonstrated the extended use in the literature of NNs for these purposes with excellent results. Therefore, a MLP model was built to detect attacks in IoT networks, where the use of constrained devices is very extended and their computing resources are limited; thus, being the perfect scenario to evaluate the capabilities of eBPF to integrate ML algorithms in commodity hardware and intelligently inspect traffic to react in real-time to cyberattacks. In order to run in the eBPF environment, the trained MLP model implementation had to be tuned to be accepted by the strict eBPF verifier before its injection in the Linux Kernel (**Objective 4**). This is a complex task given the severe restrictions posed by such kernel's verifier. Besides, for comparison purposes, another version of this network function was developed where the MLP was executed in the user plane, outside the eBPF's efficient execution environment within the kernel space. After the conducted performance evaluation, the solution implementing the data plane processing and the ML algorithm in the Linux kernel improved the CPU consumption and drastically reduced the time needed to run the MLP model. In consequence, it was demonstrated the capacities of eBPF to flexibly operate in the far edge of the network, on top of constrained devices, bringing intelligent decision making in points of the network never seen before [117] (**Objective 5**).

Finally, as the last contribution of the thesis, all the acquired expertise and knowledge in B5G network infrastructures and ZSM-based architectures was put into practice to envision how the services of the future will shape the design of 6G infrastructures [118]. The automotive vertical was taken as a basis, as it demands challenging network and computing requirements to support the upcoming wave of vehicular services. In this way, the applications envisioned to demand network capabilities still not reachable, or even not designed, were distinguished and placed in the state of the art (**Objective 7**). From this study, a set of measurable KPIs were identified. They were presented as a list of indicators that can be used to determine the performance of the network infrastructure in an objective way (**Objective 6**). The KPIs were classified depending on the segment of the network were they are illustrative to evaluate if the developed infrastructure is fulfilling the demands of the aforementioned envisioned services. Also, the study performed during all the research stages of this PhD thesis with regard to ZSM and B5G architectures was put into practice by designing a conceptual 6G platform to manage in-network applications. This platform is based on ZSM principles and experimental perspectives, where the whole lifecycle of the deployed services should be automatically monitored and evaluated to assess its performance. This is of prominent importance for vehicular scenarios in B5G networks, as critical services are a crucial part of these systems and they have to be under a fine-grain control only achievable through the use of this kind of platforms.

It is relevant to highlight that the outcomes of this PhD thesis, including the implementation of the ML-powered eBPF solution for efficient packet-processing, are being exploited and reused in European research projects. One example is RIGOROUS[9] (grant agreement ID: 101095933), where the aim is to introduce AI-based security mechanisms to dynamically react to threats in the orchestration layer and in the network functions. Also in the project HORSE[10] (grant agreement ID: 101096342), which leverages this kind of intelligent network functions to enable predictive cyberattacks detection, programmable networking, and smart orchestration and management.

[9] https://cordis.europa.eu/project/id/101095933
[10] https://cordis.europa.eu/project/id/101096342

Although the results presented in this PhD thesis can serve as a baseline for the development of lightweight AI-based traffic processing tools for B5G network infrastructures, there are still multiple research lines to be exploited to continue advancing the state of the art. In first place, the use of eBPF for implementing intelligent network functions is still in its initial stages, and the maximum potential of this synergy is still to be reached. A deeper understanding of the technology and further development efforts will create more efficient and powerful communication infrastructures capable of coping with more demanding application and service requirements. Besides, the main drawback of the eBPF environment is the limitations imposed by the verifier to assure the correct execution and termination of programs in the privileged context. Thus, by developing more optimized algorithms in a automated way, it will be possible to squeeze its performance and implantation to the next level. Likewise, following this research line will lead to the possibility to integrate more ML models, enabling the creation of a diverse range of intelligent network functions that can be flexibly run at any point of the infrastructure. By doing so, it will be possible to provide intelligence and autonomy to the whole network and computation continuum and increasing the overall performance of 6G network infrastructures. In this line, the new wave of LLMs models will have a prominent importance in this process, as they enable the quick acquisition of a vast amount of knowledge, e.g., protocol specifications, best practices, standards, existing network devices configurations, etc. This goes beyond the skills of humans and current network management systems. The integration of LLMs in the networking infrastructures will permit the unification of network intelligence through NLP to fuel intent-based network managing and orchestration, closing the gap between network administrators and the underlying network infrastructure configuration.

Finally, another research line of interest would be the integration of these network functions within the proposed 6G framework to close the loop and obtain a fully functional ZSM network architecture. This will permit the construction of more complex services across different domains using interoperability patterns. Besides, the network functions could be integrated within the presented testing and validation procedures to assess their operation after instantiation and decide whether the demanded network requirements are fulfilled. The proposed platform also monitors the provided QoS in the 6G infrastructure and can react to performance decreases. As a response, it may trigger computer and network resources reconfiguration, or the application of migration procedures in real-time to a more suitable location along the computing and networking continuum.

CHAPTER 3

Introduction

Future networks will be essential for the functioning of nearly every aspect of life and industry in the years to come, addressing the communication needs of both humans and intelligent machines. With the growing expansion of 5G starting back around 2020, more and more devices are getting connected, providing enhanced communications capabilities and shifting towards a world where everything will be connected [12]. In the coming years, society will have to rely on the network infrastructures to deliver essential and critical services anywhere and anytime, while the integrity and privacy of the data are guaranteed.

Besides, in the light of the impact climate change is having in our daily life, sustainability has to be a transversal priority. All sectors of society must address the SDGs[1] adopted by the United Nations (UN) in 2015 aiming to protect the planet, end poverty and ensure peace and prosperity by 2030. The UN identified 17 SDGs to develop social, economic and environmental sustainability. Concretely, the ninth SDG is targeted at industry, innovation and infrastructure, which is crucial for economic growth and development. It promotes sustainable industries, and investing in innovation and scientific research. Finally, it pursues the reduction of CAPEX needed in network infrastructures to close the gap between society and the digital world to ensure equal access to information and knowledge. In this line, wireless networks are already playing a crucial role to address these issues. They are an important tool to improve the resource-usage efficiency and they advocate for innovative lifestyles, which empowers a positive change in society.

Expanding network infrastructures to achieve full global coverage and close the digital gap in remote areas is essential, especially as the number of devices embedded throughout society increases dramatically. Keeping costs sustainable for both users and service providers is a critical part of this effort. Energy efficiency, a key focus in the development of 5G networks, will be even more important for NGNs solutions [13]. The expected grow in traffic should not result in a proportional increase in energy consumption, and energy consumption should be minimal when there is no traffic crossing a given network element.

To serve as the main pillar for a vast range of new and evolving services, NGNs must be enhanced and extended across various dimensions compared to today's networks. This considers not only classic capabilities, such as data rate, latency, and system capacity, but also new KPIs. It is important to ensure that future networks can support currently envisioned use cases and enable unforeseen future services. Starting with classic capabilities, future networks should provide higher data rates and lower latency across all the computing and networking continuum. This includes achieving several

[1] https://sdgs.un.org/goals

hundred gigabits per second and end-to-end sub-millisecond latency in certain scenarios. In this line, mechanisms ensuring dynamic network deployment will be crucial for the cost-effective establishment of high-capacity and resilient networks in the future. This will enhance the agility of service providers in addressing new business opportunities and emerging use cases. The main challenge is to seamlessly integrate traditional service provider-deployed network nodes with constrained, user-deployed, and mobile nodes. A common factor in all future deployment scenarios is the need for a superior transport network that is flexible, scalable, and reliable to support demanding 6G use cases and novel deployment and configuration options. This can be achieved through AI-powered programmability enabled by the softwarization of the network, virtualization over heterogeneous networks, and closed-loop automation to maintain network flexibility and manageability.

Previous generations of network infrastructures have relied on specific network equipment ruled by complex and dedicated configurations, leading to a key limitation: The inability to apply new features to legacy devices, which hinders development speed. To address this limitation, traffic handling devices should acquire programmability capabilities by replacing hardcoded behaviors with a flexible, lightweight and portable environment. This approach makes devices more future-proof and capable of supporting B5G network functionalities. Besides, it also accelerates feature development, reduces time to market, facilitates bug fixing, and promotes more agile operations. To achieve future networks capable of handling a multitude of versatile services without escalating costs and complexity, it is imperative to elevate the level of network intelligence. The resultant autonomous networks will enhance energy efficiency, optimize performance, and guarantee service availability. This transformation is anticipated to unfold in two primary ways: Firstly, through optimizations that are challenging to realize with traditional algorithms, where ML arises as a key enabler. Secondly, by evolving operational systems to autonomously manage most of system management tasks, wherein ZSM will play a pivotal role. Such an approach also involves the increasing autonomy of the system. An autonomous scheme needs innate capabilities to adapt to its environment, continuously observing and learning from past actions. With this approach, insights from operations and service performance are promptly fed back, either in short cycles or near real-time, to enhance configurations, processes, and software artifacts. Within the network logic, there will be an ongoing enhancement in AI-based algorithms, guiding real-time decision-making across physical locations and logical functions. This continuous optimization will make the system significantly more dynamic compared to current network infrastructures. Intelligence, in various forms, will be accessible throughout the whole computing and networking continuum, even in the closest points to the users.

It is also necessary to study specific verticals to understand the real requirements that the applications and services will demand from NGNs infrastructures. The vehicular environment is gaining attention during the last years as we are moving towards a fully connected society in which vehicles will have a crucial role. The World Health Organization reports that nearly 1,35 million people are killed in road traffic collisions worldwide each year, with almost 3.700 deaths occurring daily [11]. The goal of data transfer between vehicles in future networks is to disseminate information and alerts in a cost-effective and timely manner using the infrastructure. Thanks to V2X technologies, vehicles can communicate to each other as well as with pedestrians and the roadside infrastructure to exchange real-time traffic information, such as road conditions or traffic status. This enables safer and more efficient driving, reduces traffic accidents, improves road utilization, and alleviates traffic congestion. Consequently, V2V-V2I-V2P integrated networks are a cornerstone for the development of B5G infrastructures.

In the light of the previous discussion, this PhD thesis is focused on providing novel ML-based network reprogrammability solutions in NGNs infrastructures as enablers of the ZSM paradigm. To this aim, it has been researched, designed, implemented, and validated the synergies between AI models and data plane management techniques in commodity hardware. Besides, the acquired expertise was exploited to envision the integration of future vehicular services and applications within 6G infrastructures. By doing so, it was possible to address the needs of the future computing and networking continuum in B5G scenarios.

The aim of this chapter is to provide a contextualized vision of the research performed during the

development of this thesis. It reviews the research works in the literature and the different identified gaps that motivated the addressed research lines. The remainder of the chapter is organized as follows. Section 3.1 summarizes the current challenges for the autonomous management and orchestration of future network infrastructures. Then, Section 3.2 analyzes the state of the art of the relevant paradigms and technologies considered in this thesis, as well as the identified gaps. Finally, Section 3.3 states the lessons learned and conclusions from this research journey.

3.1. Autonomous management and orchestration of future network infrastructures: Challenges and limitations

The autonomous management and orchestration of network infrastructures is currently a hot topic in the research community [14–17]. It will revolutionize how network infrastructures are operated and maintained. The desired autonomy will allow real-time governance with immediate reaction to changes in the network or in the user's behavior. To achieve this ambitious objectives, it is necessary to shake the current network paradigms and provide unified solutions that can handle every aspect of the network management and orchestration.

Recent research efforts [14] has analyzed the main challenges that this paradigm shift will pose to the community. The authors of this work conclude that the evolution towards completely self-managed networks needs the coordination of all the segments of the infrastructure, from the physical devices to the top-level orchestration components. This is the reason why the envisioned challenges are distributed along the whole networking stack. These challenges mainly emerge from the necessity of handling future services and applications, which are demanding ubiquitous computing capabilities and a formidable network performance supporting high bandwidths and close to zero latencies. Besides, the introduction of intelligence within the infrastructure governance defies current architectures and management hierarchies.

In this way, the following subsections summarize and discuss the faced challenges to provide autonomous management and orchestration for future network infrastructures.

3.1.1. AI limitations

AI technologies, specifically ML, are envisioned as the cornerstones for the implementation of completely automated networks. AI models enable the desired self-managing capabilities, improving service delivery and significantly reducing the operating expenses. However, the integration of AI techniques within autonomous management and orchestration procedures faces different limitations and risks.

The main limitation to introduce ML reasoning in the network is the lack of high-quality datasets. It is crucial to train the AI models with appropriate data, as their accuracy depends on it. 5G traffic datasets are expected to be essential to develop efficient algorithms for the automation of the network. However, there is a limited set of that kind of datasets, as the 5G networks are still in their rolling out process and the operators have to go through a complex privacy preserving process in order to release them. These datasets should contain complete, accurate and timely data to produce relevant AI models to optimize the intelligence of the decision-making process in the management and orchestration of the network. Besides, the size of the desired datasets is huge, and they should be labelled, which means that the operating costs are high, as well as the computer resources needed to train the ML algorithms. For these reasons, the required training to successfully make AI techniques resolve complex problems with great accuracy takes a lot of time and a considerable amount of resources, which makes it difficult for them to be used in real-time use cases. In addition, in the heterogeneous B5G environments, data patterns will change continuously due to the nature of the services and the ever-changing behavior of the users. This requires the models to be constantly retrained and limits their use in online scenarios. In consequence, the maintenance of high accuracy models while reducing the training time remains as a crucial challenge to provide intelligent governance to NGNs. Also, the reduction in the computing

resources usage, both in terms of CPU and memory, together with the use of less energy-demanding processes are important points to consider for more sustainable networking processes.

3.1.2. Scalability

Network infrastructures are traditionally geographically distributed using a wide range of traffic management technologies and optical backbone links. The inclusion of automated decision-making in them requires a deep knowledge of the status of the network throughout all the physical infrastructure. There are two ways to address this issue, namely, centralized or distributed. By using the first one, the central entity in charge of management and orchestration tasks will collect data from the whole network and its decisions will be based on a complete vision of the system, thus achieving high accuracy by holistically optimizing the behavior and performance of the infrastructure. However, it depends on the correct and fast gathering of information from every segment of the network, which may be costly in terms of generated traffic, overloading the infrastructure, and the latency to receive crucial data may be high. On the other hand, distributed approaches benefit from local inference processes to take quick actions in more limited scenarios, therefore decreasing the overall latency to make decisions. Nevertheless, its accuracy can be degraded as the elements making decisions lack a complete view of the infrastructure, thus it may lead to incorrect actions or disalignments. Besides, if hierarchical methods are implemented, the error accumulation may also decrease the performance of the system.

3.1.3. Ethics

To automate the management and orchestration of network infrastructures, AI-powered mechanisms need to comprehensively analyze the traffic running through the network, as well as to study and predict the behavior of the users. The gathered data will then be processed and they will fuel the decision-making process to drive the configuration of the network and the usage of the resources. This raises significant ethical concerns with regard to the privacy and security of the collected data, as well as possible manipulation of data streams or the injection of malicious or fake information in the AI models. Besides, the explainability of the made decisions is important to make humans aware of the behavior of the system and the automated governance resolutions. It has to be understandable how the attained outcome is generated and the reasoning behind the processing of the input data. Finally, trustworthiness is a fundamental pillar of the automation of orchestration and management, as the AI decisions have to be coherent and trusted by the humans designing and overseeing the system.

3.1.4. Security

The embracing of ML-based automation in the network infrastructures widens the horizon for new attack vectors and requires careful study and attention from the security viewpoint. The objective of eliminating humans from all network governance procedures aims at reducing the operational costs and human errors, as well as improving scalability and increasing the overall performance of the system. However, certain security considerations cannot be overlooked. ML algorithms can be disrupted in multiple ways, causing malfunctions in the whole system. The collected data fueling the different AI models can also be altered by malicious actors, affecting its integrity and impairing the accuracy of the predictions. Availability is another key aspect in these envisioned autonomous frameworks, as it is expected the reduction of the human workforce supervising the systems. Well-targeted attacks can interrupt key autonomous modules and stop the operation of the network. Thus, measures should be introduced to minimize the attack surface and the downtime in these events, providing self-healing capabilities. Finally, additional security mechanisms should be introduced to replace the typical monitoring activities of hardware devices and the overall system made by technicians and engineers.

3.1.5. Hardware investment

As things stand today, the computational and networking resources needed to introduce the required intelligence and processing levels for the envisioned NGNs will imply a colossal investment. The expected amounts of traffic crossing the network, with the required data for network automation, will need either expensive specific equipment or a significant advance in the packet processing mechanisms in network devices. Although the core part of the infrastructures may be considered ready for the first stages of the evolution, from this domain to the final users there is still a gap that have to be addressed, particularly in the radio access network segment. In consequence, it becomes of utmost importance to develop high-performance software techniques for packet handling tasks able to run in cost-effective equipment easily located at any point of the network infrastructure. Besides the efforts in the networking segment, the computing resources needed for the foreseen volume of AI-related tasks are huge. B5G networks are expected to automate all the decision-making process, which implies ubiquitous processing across the whole network continuum. This is translated to the deployment of a lot of equipment with high computational capacities everywhere. But this approach will not be enough by itself; the processes and inference activities should be optimized to the maximum level to reduce not only the CAPEX but also the energy consumption, as the pursued model should be sustainable and aligned with the needs of the society.

3.1.6. Services life-cycle management

In such complex environments as the ones that will be found in NGNs infrastructures, managing the life-cycle of the services running on them is essential. With the automation of the network, it is crucial for the developers to make sure that the applications deployed over the infrastructure are compatible with this new paradigm and that it will not suffer any service disruption. From the operator standpoint, it needs to be sure that the third-party applications being deployed in its premises will correctly function and that they do not pose any threat to the network. In this way, it is fundamental to come with B5G frameworks capable of completely managing the life-cycle of services and applications. These platforms should have an experimental approach, which enables the validation of the proper operation of the deployed services by checking its compliance with already defined QoS network requirements, fulfilling concrete KPIs. Besides, they may also include continuous infrastructure monitoring, perfectly aligned with the requirements of AI-based network infrastructure management and orchestration.

3.1.7. End-to-end management

In order to support the requirements of future services and applications, it is necessary to effectively manage the end-to-end path, from users to applications. Traditional automation frameworks lack generality and adaptability to handle the heterogeneity present in B5G infrastructures, which is even harder when talking about end-to-end services crossing different domains. In these cases, it is common to find that each segment of the network is owned and maintained by different stakeholders. In consequence, it is challenging to provide automated high level management and orchestration, as each one of the infrastructure owners may have different policies, equipment, and network requirements. Thus, it is crucial to articulate the needs of the network with a business-level model that enables multi-domain governance. This model can be understandable for every domain manager so it can enforce the desired configurations and requirements in its domain. This issue is closely related to all the already discussed challenges, as it implies addressing all of them in each domain, but also considering the problem from an holistic perspective to reach and adequate coordination in an end-to-end fashion.

The autonomous management and orchestration of future network infrastructures is rapidly gaining prominence as a critical area of focus for improving and integrating into real-world NGNs deployments. The surge in research proposals and standardization efforts exploring it from diverse perspectives, such as resource orchestration, traffic monitoring, and cybersecurity, underscores its growing significance.

However, the integration of ML mechanisms into infrastructure management systems introduces a plethora of challenges. These challenges primarily revolve around AI limitations, scalability, ethics, security, hardware investment, services life-cycle management, and end-to-end management, which can pose significant obstacles to achieving the desired levels of QoE for users. Furthermore, the continuous evolution of ML, NGNs, and the applications and services they enable needs a dynamic approach to autonomous management and orchestration development. Novel architectural designs, innovative virtualization schemes, and the development of advanced management functions are crucial for enriching the already vibrant and expanding B5G ecosystem in the foreseeable future. This ongoing revolution is essential to ensure that network governance keeps pace with the rapid advancements in technology and continues to deliver optimal user experiences. To successfully integrate into society and create a positive impact, the new wave of services and infrastructures must address the outlined challenges. This involves developing solutions that bridge the digital divide, while prioritizing energy efficiency, data privacy, and ethical considerations. In the following, a comprehensive review of the research efforts paving the way for achieving such ambitious goals is presented.

3.2. Related work

This section presents the foundational related work for the development of this thesis, covering all the key topics that compose its main pillars. These areas are: (i) ZSM, (ii) data plane reprogrammability, (iii) integration of AI in data plane management, and (iv) service-driven platforms for B5G infrastructures.

3.2.1. Zero-touch Network and Service Management

The rolling out of 5G infrastructures with initial network slicing capabilities has boosted a paradigm change in the management and orchestration of networks and services. In consequence, it arises the necessity of addressing (i) the transformation of traditional network architecture models into programmable, software-driven, service-based and holistically managed architectures, and (ii) the operational flexibility needed to support new business models enabled by disruptive networking technologies, e.g., network slicing, AI mechanisms, data plane reprogrammability, etc.

These deployments bring together a wide range of stringent network requirements to achieve near to zero latency, ultra-high reliability, huge volumes of traffic, personalized services, and support for massive densities of devices. To fulfill these objectives, it is essential to achieve full end-to-end automation of network and service management, while ensuring the economic viability of the offered services. The ultimate goal is to create big autonomous networks controlled by high-level policies and rules. This will provide the network infrastructures with autonomous innovative capabilities such as self-configuration, self-monitoring, self-optimization, and self-healing. To achieve this ambitious goal, it is necessary to come up with an end-to-end framework capable of leverage AI algorithms and models to realize closed-loop automation.

The ETSI took the lead in this task and founded the ETSI ZSM working group in December 2017 to accelerate the design and definition of the envisioned end-to-end framework and the underlying technologies enabling it [2]. The group is dedicated to establish collaboration relationships with relevant standardization bodies, open-source projects, and industry groups to promote the adoption and alignment of the ZSM architecture and solutions. End-to-end automation is a significant undertaking and represents the industry's focus for the coming years. The implementation of AI models within network management and orchestration frameworks will evolve gradually, with insights from real deployments and operational experience informing the specification work. Thus, the ZSM working group encourages the development of proof of concepts to demonstrate the practicality of ZSM implementations. The outcomes and lessons will be integrated into the group specification work. Additionally, it will consider feedback and findings from actual deployments and operational experiences in its specifications.

[2] https://www.etsi.org/technologies/zero-touch-network-service-management

The ZSM reference architecture [18] evolves from typical stiff management services towards a more flexible design. The architecture is divided in different building blocks to enable the construction of complex service chains in a modular way, as it can be seen in Figure 3.1. It is composed of management domains containing distributed management and data services, which are integrated via an integration fabric that enables the intercommunication, service consumption and connection with third parties. On top of them, there is a cross-domain integration fabric, whose aim is to interconnect them to provide end-to-end capabilities through the exposure or consumption of service end-points, enabling full ZSM capabilities. This design permits the new services and modules to be accommodated independently so they can be deployed in an independent way. By doing so, it enables reusability, portability and the inclusion of vendor-neutral resources.

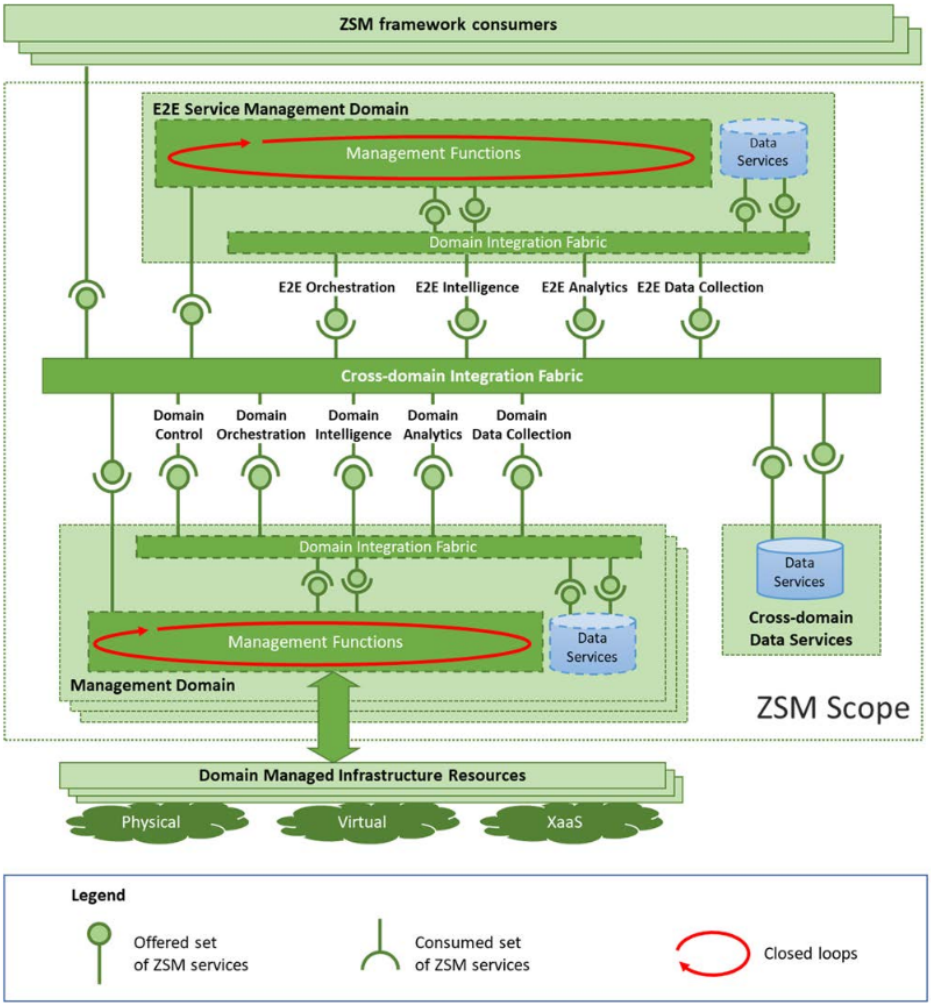

Figure 3.1: ETSI ZSM reference architecture. Extracted from [18].

The main component of the ZSM framework is the management service, which has certain capabilities that can be exposed through standardized end-points to be consumed. These capabilities describe the function of the management service within the organization it belongs. To provide modularity, the management services can be chained or merged to create new ones and expand the capabilities of the system. The resources of the infrastructure, either physical or virtual, can communicate to the management services via their end-points to inform about their specifications and their features.

The intelligence services of each domain [18] implement the closed-loop management and orchestration automation by supporting different levels of decision-making inferences. The offered services can be classified into three different classes, namely, action planning, decision making, and decision support. This last kind of services are the ones implementing AI models at the lowest level to enable the realization of the decision making at a higher level. The information fueling this process is gathered by the already defined ZSM services located in the analytics and data collection domains. On top of them, taking as an input the considered decisions, the action planning services establish the management and orchestration actions to be carried out by the control and intelligence domains.

Other relevant standards

Although the ETSI ZSM specification is the reference framework when considering the automation of network management and orchestration, there are other relevant standardization efforts pushing towards the same goal promoted by multiple organizations. Actually, ETSI also created another group to develop new specifications for cognitive management systems: the ETSI ISG ENI [3]. Its aim is to define a cognitive network architecture rooted in the use of AI models and context-aware policies to detect changes in user needs, environmental conditions or business goals and adjust the offered services accordingly. Therefore, it seeks to provide automated service provision, operation, and assurance, while optimizing the resource orchestration and the slice management. In addition, it also assists the decision-making of the humans to enable more maintainable and reliable systems. During the last years, the group has produced several specifications that are already published and available online, e.g., the system architecture [19], the terminology [20], the requirements of the system [21], the envisioned use cases [22], and the definition of data processing mechanisms [23].

This is accomplished through policy-driven closed control loops that leverage emerging technologies such as big data analysis and ML. These technologies adjust the configuration and monitoring of networks and networking applications. The system dynamically updates its knowledge base to understand the environment, including the needs of end-users and the goals of the operator, by learning from its own actions and those taken by humans, making it an "experiential" architecture. It also ensures that the automated decisions made by the system are accurate, enhancing the reliability, stability, and reducing the maintenance efforts. Additionally, it can determine the appropriate services to be considered in an certain context, and which ones are in risk of not fulfilling their corresponding SLAs. This process is supported by the telemetry collection mechanisms of the architecture, which are later used to assist in the monitoring of the system and to later optimize the infrastructure performance.

Figure 3.2 shows a high-level overview of the ENI cognitive architecture with an integrated API broker. Its pillars are the input processing, the analysis of the data, and the output rendering. The broker enables interconnection with other external systems that are not compliant with the ENI definitions and interfaces. It also permits the processing of data gathered from the infrastructure that is not compliant with ENI formats. In the same way, it can emit recommendations or commands in alien formats not understandable within the ENI system.

Another significant effort by ETSI is made with the GANA model, which was defined by the Autonomic Management and Control intelligence for Self-managed Fixed and Mobile Integrated Networks Working Group. The GANA reference model [24] provides autonomic communications, networking, and cognitive management and control. The foundational block of this design is the physical or virtual resources that can be managed by the decision element, which monitors them to compare their status with the desired one. If it changes, the decision element creates a plan of actions to dynamically correct the behavior and come back to the desired status. In a higher level, the knowledge plane directs the different decision elements to achieve different objectives, such as configuration modifications, QoS, or QoE. At this point, the knowledge plan maintains a high-level view of the network infrastructure so it can control the whole domain. This model is designed as an abstract approach to self-managing capabilities in NGNs and it is not limited by any specific technology or implementation-oriented architectures.

[3] https://www.etsi.org/technologies/experiential-networked-intelligence

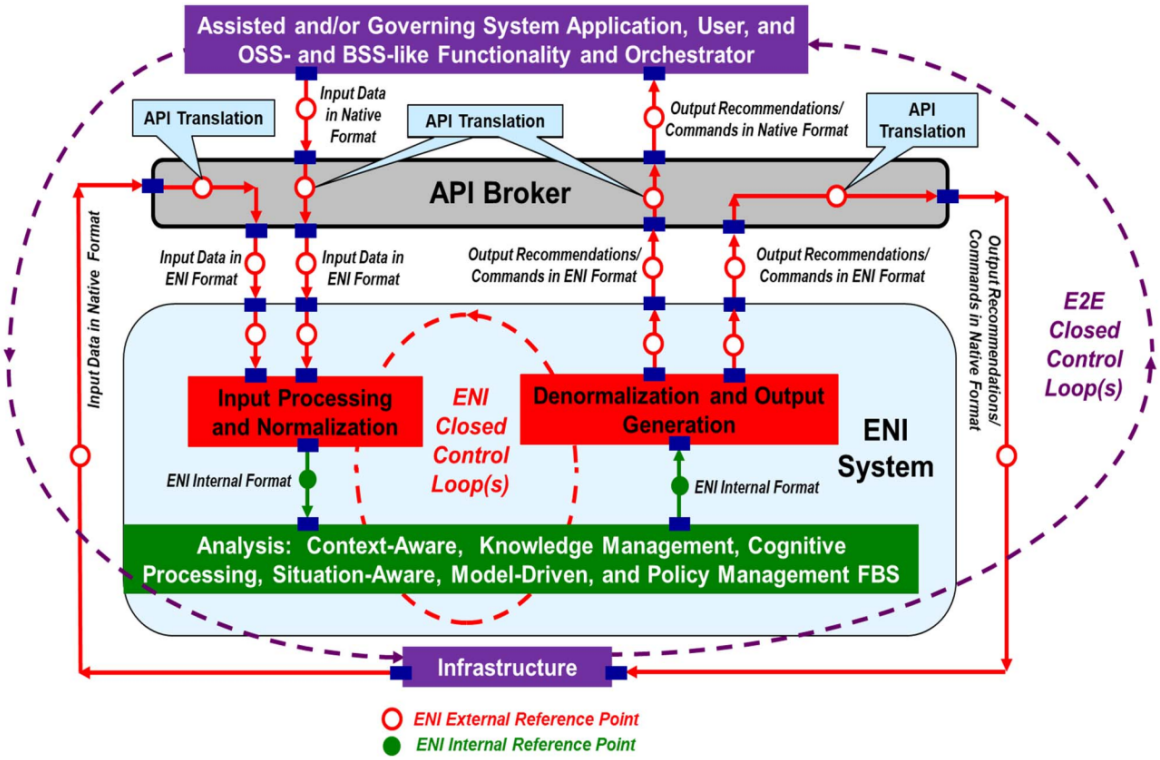

Figure 3.2: ETSI ISG ENI reference architecture. Extracted from [19].

The ITU also contributed to the definition of network automation architectures with the ITU-T Focus Group on ML for Future Networks. It was active from 2018 until 2020 and its objective was to draft technical specifications for the use of ML techniques in NGNs. It was an open initiative in which both members and non-members could collaborate to explore the application of ML in B5G networks. The working group managed to publish several recommendations, such as an architectural framework for the integration of ML in future networks [25], a proposal of a framework for a ML-based marketplace for NGNs [29], and it also evaluated the possible use cases [26], the intelligence levels [27], and data handling procedures [28].

As a fundamental element of the proposed architectural framework, the ITU-T working group defined a ML-based pipeline for the management of NGNs infrastructures. As it can be seen in Figure 3.3, the framework is based on three components. The ML pipeline is a chain of functional nodes that can be combined to create an intelligent network function. On top of it, the MLfunction orchestrator directs and coordinates the nodes in the pipeline by selecting the appropriate ML models and sorting the nodes. At the highest level, the ML sandbox acts as a digital twin of the deployed pipelines to train, test, and evaluate them before their deployment in production environments.

In a similar way, the 3GPP Service and System Aspects Working Group 2 also proposed a framework for data collection and analytics exposure to enable network automation in 5G infrastructures [30]. This framework interacts with different entities for multiple purposes. To provide data collection capabilities, it gathers data from functions in the 5G core, and information from untrusted sources via the NEF. In the same way, the framework exposes the collected data to other network functions within the 5G core, and towards untrusted application functions through the NEF. The standard also contemplates the acquisition of ML model information on the related gathered analytics. Additionaly, through a similar procedure, it permits the obtention of network slice load level analytics, which is calculated by the framework based on the information collected from the different components of the

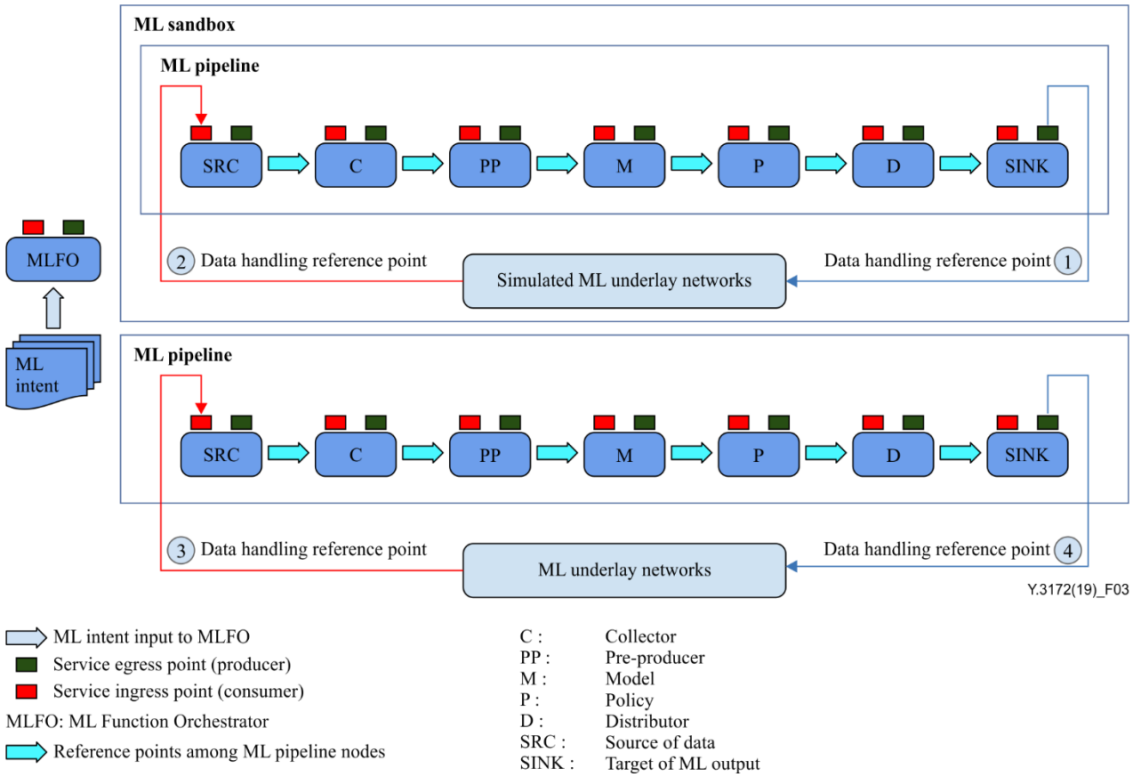

Figure 3.3: ITU-T reference architectural framework. Extracted from [25].

5G core. These capabilities set the foundations for the adoption of autonomous management and orchestration mechanisms in B5G networks.

Besides, the 3GPP was also a pioneer in this field by proposing back in Release 8 of its wireless broadband communication standard the concept of SON. They considered that the network operators were going to need new capabilities from network infrastructures, such as management flexibility to reduce the operating costs. 3GPP's proposal has evolved since its initial definition [31] until Release 18 [32], although it was decided that the SON algorithms themselves were not going to be standardized. It is important to highlight that the proposed 3GPP systems using SON do not rely on AI mechanisms for decision-making. Instead, they implement a closed-loop paradigm to provide autonomous capabilities. Therefore, the SON algorithms may consist of monitoring agents to collect management data, analysis elements to determine if there are issues to be resolved, a decision engine to take the resolution, an execution component to run the decided actions, and an evaluation module to asses the success of the applied action. The specification defines three types of SONs based on the location of the algorithms: (i) Centralized, in which the SON algorithms are executed in the 3GPP management system (either in the same domain or in cross-domain scenarios); (ii) distributed, in which the algorithm is located in the networking functions themselves; and (iii) hybrid, where the SON algorithm is executed in both levels.

Several of these standardization bodies are actively exploring the application of ML for ZSM. ETSI ZSM and ETSI ISG ENI are particularly focused on applying AI for network automation and orchestration. ETSI GANA contributes with a reference model for ZSM network architecture design. While ITU-T also explores ML for infrastructure management, their vision does not encompass fully automated architectures. Interestingly, 3GPP utilizes a closed-loop architecture for network optimization, where elements are monitored and algorithms react to changes, though not through

ML techniques. These initiatives are in their infancy stages, with their primary contributions to date being reference architectures. We can expect more specific advancements in the future, so ZSM and AI researchers should stay informed on updates from these organisms as the proposed architectures evolve and gain new functionalities.

ZSM-enabling network management and orchestration functions

In the following, research proposals using AI to develop network functions enabling ZSM in NGNs are explored. The categorization of the reviewed solutions is based on their role within the network and the networking operation they automate using ML. In this case, it is considered flow inspection, multi-domain management, RAN management, and network resources management. This review provides a concise overview of recent ZSM-related research, exploring various approaches and methodologies. An extensive search in major academic databases like Science.gov, Google Scholar, Microsoft Academic, and Semantic Scholar, along with repositories from technical publishers, has been conducted.

As the first category, ZSM systems rely on **traffic inspection** to categorize each data flow. This involves analyzing its origin, destination, data type, and priority. By understanding these characteristics, ZSM can automatically route traffic efficiently and ensure it receives the appropriate level of QoS. Authors in [33] presented a network slicing architecture that leverages ML for early mobile app traffic classification. By analyzing the first five packets of each flow and source/destination ports (pre-processed for consistency), they grouped applications with similar traffic patterns using K-means clustering. This created three categories, which then served as training data for different ML algorithms to classify future traffic. Notably, all algorithms achieved high accuracy (>96%), with GBT and RF excelling a near-perfect (100%) classification. In [34], it was introduced a deep learning approach for traffic classification within SDNs. This method prioritized application-specific QoS by analyzing packet payloads. The classifiers resided in the SDN's control plane, influencing data plane routing decisions. To enhance model adaptability to unseen traffic, they focused solely on payload data, excluding packet header information. The payload was converted into image data by grouping bits into pixels. Two deep learning models were evaluated: A multi-layer LSTM network and a combination of a single-layer LSTM with a LSTM. Hyperparameter tuning optimized each model for the specific dataset. Experimentation revealed that the multi-layer LSTM model outperformed the combined approach, suggesting its potential as a robust solution for network traffic classification.

Phan *et al.* [35] proposed a RL-based framework to optimize traffic flow matching in SDNs. This helped to prevent flow-table overflow in SDN switches and improved monitoring performance. The approach used Q-learning to adjust traffic granularity in the data plane, treating the network devices as the RL environment. Additionally, a policy creation module within the Q-DATA framework analyzed traffic using SVMs to predict performance degradation. Based on this information, the module determined the optimal action for improved traffic flow matching. Real-world experiments showed significant performance gains compared to traditional SDN controllers. In turn, work in [36] proposed a ML approach to predict traffic demands in optical networks with chained VNFs. Authors modeled the network as a series of nodes and connections, with traffic represented by flows between source and destination points over time. Their goal was to predict which source and destination pairs would experience traffic in the next time interval. To simplify the problem, they transformed it from multi-class (predicting all possible traffic patterns) to binary classification (predicting whether any traffic occurs between each pair). Eight different classification algorithms were evaluated, finding that Linear Discriminant Analysis performed the best overall. Interestingly, the length of VNF chains significantly impacted performance. Longer chains generated more possible traffic patterns, leading to more accurate predictions

As it can be seen from the reviewed works, traffic flow classification struggles with limited training data, but the explored proposals tackled this with a two-stage approach. In the first stage, authors of [33, 35] used clustering and SVMs to create training data and then applied ML [33] or RL [35] for classification. Others, like Lim *et al.* [34], used the same data for both stages but separated header information for training and payload for classification. All studies highlighted GANs as a promising

solution to generate realistic artificial training data. Notably, SL emerged as the preferred approach for traffic flow classification due to existing SLAs and network policies that guide classification and QoS actions.

The next relevant role in which network functions can be classified in ZSM environments is the **multi-domain management**. With the growing trend of physical resource sharing between network operators and the breakdown of network functions into independent and modular entities, the effective management of these distributed network components becomes essential to ensures the smooth operation of the entire system. Several works found in the literature aim at addressing this issue with different approaches, as discussed below.

Chen *et al.* [37] addressed the challenge of managing complex multi-domain optical networks by proposing a cognitive framework with multi-agent learning. The framework utilized intelligent broker agents equipped with DNNs. These agents analyzed network data to infer optimal service provisioning strategies. When a service request arrived, the broker analyzed the current network state and recommended an action to the service manager using DRL. The learning process was further enhanced with feedback from the domain manager, allowing the agents to learn optimal policies quickly. The study demonstrated that the approach outperforms other schemes by achieving higher rewards. In [38], authors addressed inter-domain service provisioning in elastic optical networks with the design of a hierarchical learning framework. This framework used a two-layered approach: Domain managers handled local networks, providing services, monitoring traffic, and predicting intra-domain transmission quality with cognition agents, namely, NNs. Meanwhile, brokers managed inter-domain requests and performed global optimizations. To establish inter-domain services, the brokers leveraged information from domain managers, including available paths and predicted quality. While experiments showed efficient provisioning, the framework's scalability required further investigation and extensive validations.

Work in [39] tackled inter and intra-domain light path provisioning with a ML technique for estimating Quality-of-Transmission. The solution utilized a NN with two hidden layers (each with 10 neurons) to analyze power measurements and noise levels across links. This configuration achieved the optimal balance between accuracy and complexity. Experiments showed an average estimation error of less than 6%, demonstrating the effectiveness of the proposed approach. Authors of [40] proposed a ML-based approach for dynamic spectrum sharing in cellular networks with multiple operators. Their scheme utilized RL to fairly allocate channels and minimize interference issues. Mobile network operators learned optimal radio resource allocation through a reward system based on Signal-to-interference-plus-noise ratio, where positive rewards were given for high values, indicating successful channel access, while negative rewards penalized interference events. Simulations with two operators demonstrated the scheme's effectiveness in achieving fair spectrum sharing and increased network throughput.

Thar *et al.* [41] came up with a ML solution to improve profit for mobile virtual network operators through efficient resource management. Their scheme used DL to predict video content popularity, allowing them to share virtual cache storage at base stations. This permitted to reduce backhaul network usage and user access delays. To find the optimal DL model, e.g., CNNs,RNNs, or a combination of them, an RL searching scheme guides exploration towards high-performing models. A central controller located popular content, trained the model, and sent recommendations to the base stations for storage. The conducted experiments validated the efficiency and accuracy of the generated models, demonstrating significant network traffic reduction through content caching. The proposal in [42] tackled the challenge of synchronizing controllers in multi-domain SDNs by exploiting a RL approach with DNNs. Authors modelled the problem as a Markov decision process and trained a policy called Multi-Armed Cooperative Synchronization to optimize network performance through controller synchronization. DNNs enabled these policies to learn evolving network patterns and maximizing the use of limited synchronization resources. Simulations demonstrated that this approach significantly outperforms existing SDN controller synchronization algorithms.

Relevant conclusions can be inferred after exploring the multi-domain management approaches in ZSM. While hierarchical approaches are predominant, their scalability remains uncertain. Local

agents lack full network visibility, hindering inter-domain path prediction; however, granting them global views raises privacy concerns. Additionally, error propagation through the hierarchy reduces performance (as seen in [38], [39]), requiring error correction mechanisms. Thereby, some research focuses on improving hierarchical learning efficiency [38]. The high cost of training real-time ML models with sufficient data in such architectures is another challenge. RL offers a promising solution, balancing information gathering and model selection for optimal predictions.

Another important category in the classification of AI-powered network functions enabling ZSM systems is the **RAN management**. NGNs with multiple radio access technologies will need sophisticated handling schemes to efficiently utilize radio resources. Integrating intelligence into this network segment allows for the development of smart schedulers and resource managers. These intelligent systems will deal with diverse traffic demands with high-performance techniques. The works explored hereunder show the different approaches that have been adopted to address this issues in the last years.

Albonda *et al.* [43] addressed dynamic radio resource allocation for V2V communications in NGNs with RAN slicing. Their approach used the well-known offline RL algorithm called Q-learning with soft-max decision making. To avoid potential performance issues during exploration, the RL algorithm was trained in a simulated network model before real-world deployment. The slicing controller then executed the trained policy to allocate resources for uplink and downlink traffic in each network slice. Evaluations using MATLAB demonstrated that this approach improved network performance compared to existing solutions. In [44], the authors proposed a DL approach for MIMO channel estimation that tackles inter-user interference, a common issue that increases estimation errors. The scheme used a jointly trained system with two-layer NNs for pilot design and DNNs for channel estimation. All NNs were trained offline using realistic channel and noise data to minimize the mean squared error of channel estimation. This approach allowed for non-orthogonal pilots and a non-linear channel estimator. Extensive simulations demonstrated that the proposed scheme significantly outperformed existing linear estimation methods and exhibits robustness to variations in signal-to-noise ratio.

Authors of [45] designed an intelligent channel allocation scheme for high-altitude 5G platform stations using massive MIMO systems. The scheme combined Q-learning RL with back-propagation NNs for autonomous learning. They modelled channel allocation as a MDP and leveraged RL for optimization. However, the large number of connections inherent to massive MIMO systems created a vast RL state space, making direct management cumbersome. To address this, a back-propagation NN was employed to estimate Q-values. Every Q-update served as training data for the network. Simulations demonstrated that this approach substantially improved performance when compared to random channel allocation solutions. In turn, authors in [46] developed a DL approach for joint channel estimation and pilot design that tackles fading, a signal degradation issue, in two scenarios: Quasi-static and time-varying. They leveraged GANs to generate realistic channel data for training. In the quasi-static case, a deep autoencoder with a feedforward NN and a CNN decoder learned channel coefficients and optimized pilot signals based on received signal-to-noise ratio feedback. The time-varying scenario employed a similar approach but combined a RNN and a CNN to capture temporal features more effectively, with an additional LSTM network further enhancing temporal learning. Extensive simulations and experiments demonstrated that this scheme greatly improved performance and effectiveness compared to existing methods.

Lynch *et al.* [47] proposed an automated approach using evolutionary learning to design link allocation algorithms for 5G heterogeneous wireless networks. This eliminated the need for costly manual design efforts. The evolved schedulers leveraged real-time link quality reports to optimize controller design. The system generated RAN schedules by analyzing statistical features extracted from link quality measurements and then mapped those features to optimized schedules for each cell. They evaluated two models: A grammar-based genetic programming model and a fixed-topology NN with genetically optimized weights. Simulations in a enterprise environment with 12 LTE and 8 WiFi cells demonstrated that the evolved schedulers outperformed baseline heuristics, achieving better network performance and downlink rates. Work in [48] presented a lightweight RL-based scheme for distributed channel selection in massive IoT communications. Considering the limited memory and computational capabilities of IoT devices, authors modelled the problem as a multi-armed bandit problem and

aimed at maximizing successful data transmissions. The scheme leveraged a tug-of-war strategy, a technique for maximizing rewards from RL, to explore suitable channels. It simply checked for received acknowledgement frames to determine the channel's effectiveness. This lightweight approach allowed the scheme to run on resource-constrained devices and simulations demonstrated its ability to dynamically select the best channel while maintaining fairness among devices.

Work in [49] explored a ML approach to improve channel allocation speed in cognitive radio networks for Voice-over-IP traffic. The proposed system used a two-dimensional channel distribution model based on SVMs to identify free and occupied frequency bands. This approach improved upon traditional linear models and reduced channel search time by 10% in experiments, leading to faster transmission speeds. In [50], authors developed a RL-based slice admission strategy for 5G flexible RANs, where multiple providers shared the virtualized network infrastructure. A RL agent embedded in the orchestrator learned to manage slice admission, setup, scaling, and tear-down to maximize the infrastructure provider's profit while handling services with different and varying priorities. The system modeled it as a loss minimization problem and used a NN-based agent optimized with gradient descent. Through computer simulation, this approach achieved a 50% and 23% reduction in loss compared to static and threshold-based heuristics, respectively; thus, demonstrating its effectiveness in managing slice admission for multi-service providers. Finally, Sandoval *et al.* [51] proposed a lightweight RL framework to aid IoT devices to choose the optimal radio access technology for reporting events. This framework, based on a type of genetic algorithm called evolution strategies, considered the device's overall state and prioritized maximizing performance while operating on resource-constrained devices. The RL agent received rewards that boosted faster reporting, considering factors like message priority, data packet size, and transmission delay. Measured in bits per second, the reward function prioritized efficient data transmission. Simulations in both cellular and LoRa networks demonstrated that the developed scheme significantly outperformed existing solutions, achieving a 75% improvement in the obtained rewards.

RL is at the forefront of automating RAN management. The inherent dynamism of radio environments demands fast adaptation from the ML models. RL excels at this, often aided by NN, through continuous learning from real-time feedback. However, constant online training of the model becomes computationally expensive and elevates the costs. Therefore, most research works propose the offline training of the RL solution, followed by embedding it within resource-constrained devices, essentially transforming them into smart objects [52]. Current research efforts explore new ways of use RL for tasks like scheduling, resource allocation, channel selection, and estimation.

Finally, the last categorized type of network functions enabling automated ZSM systems is the **network resource management**. Managing resources in large, decentralized networks is complex due to the distributed allocation of components and the numerous potential failure points. ML-based monitoring can address these issues by learning anomaly patterns to detect and even predict network faults. Additionally, NFV allows the migration of network functions from dedicated hardware to software running on commodity hardware. However, mapping virtual resources to physical ones presents several challenges. To address them, ML-based approaches can replace static legacy solutions, automating and streamlining the resource management processes. In the following, we discuss the latest works found in the literature researching this matter.

Building on the concept of end-to-end performance prediction, work in [53] proposed a system utilizing RL for both VNF performance forecasting and placement automation. The multi-layered architecture leveraged multiple agents: An application monitoring agent tracked application performance, a node-monitoring agent supervised resource utilization, a prediction agent anticipated VNF performance feeding information to an orchestrator, and a placement agent strategically instantiated VNFs. This RL approach employed adaptive Q-learning to predict the total service time for applications running video transcoding VNFs, considering both transmission efficiency and VNF processing power. Experiments showed that the RL-based system not only outperformed traditional SL models in VNF performance prediction accuracy by 45%, but also demonstrated superior adaptability to dynamic network traffic fluctuations. Authors of [54] designed a DRL framework to manage MEC systems without needing intricate details about the underlying infrastructure. The approach was tested in a

simulated LTE network where base stations were connected to MEC servers acting as local storage units. The core of the system is a DRL model that determined the optimal moments to transfer data between base stations according to user location and current network conditions. To achieve this, the DRL model incorporated a DNN that predicted state values, guiding the model's decision-making process. Evaluations within a simulated environment composed of a MEC-enabled LTE network and the DRL engine showed that this data transfer strategy improved the overall system performance compared to an scenario without any dynamic data trasnference.

In [56], a novel DRL-based network architecture was introduced for automatic routing in SDNs. This architecture leveraged a closed-loop system where a central SDN controller interacted with the network and a DRL agent to optimize traffic flow. The controller collected network data and translated the DRL agent's chosen action into flow table rules, essentially dictating how packets are routed. A traffic monitoring module played a crucial role by predicting traffic patterns and implementing congestion-avoidance policies generated by the DRL agent. Through this closed-loop interaction, the DRL agent learned and refined its decision-making, ultimately achieving near-optimal routing configurations. Compared to traditional routing protocols, the DRL-driven approach exhibited faster convergence and superior performance, as evidenced by reduced packet delays and increased network throughput. Work in [57] proposed a dynamic resource management system for networks that leverages RL to optimize resource allocation in scenarios with multiple tenants using the same network infrastructure. The system allowed tenants to negotiate with the provider and manage their resources to maximize their own profits. The core of this approach was a dynamic resource trading system modeled as a MDP based on Q-learning. By learning from past actions and adapting to various traffic demands, the system allocated resources efficiently. Evaluations using computer simulations showed significant improvements in tenant profit compared to existing fixed allocation methods.

In [38], the authors introduced a DRL-based system for orchestrating network slices. This system used a learning agent that dynamically allocated resources across the RAN, computing nodes, and the transport network. The agent continuously learned from network demands and optimized resource allocation to ensure end-to-end traffic flow performance meets the SLAs of each slice. This orchestration was achieved through a combination of techniques: Virtualizing radio resources with a hypervisor based on Open Air Interface that incorporated new user scheduling methods, managing traffic flow bandwidth with SDNs, and controlling computing resources by regulating the number of threads allocated per user. Evaluations demonstrated a notable improvement compared to a baseline approach. Raza *et al.* [58] investigated how to use ML to manage slice admission in 5G networks, comparing two approaches: SL and RL. The SL approach used historical network data and big data analytics to predict future resource demands for each slice. This allowed the system to reject incoming slice requests that would cause resource conflicts. The RL solution employed a stochastic policy network modeled by a NN, which analyzed the current resource allocation and the incoming slice request, and then came up with the probability of accepting or rejecting the request. The network orchestrator leveraged these data to make the final admission decision. Evaluations demonstrated that both ML approaches outperformed static admission control methods, reducing infrastructure provider losses. In turn, authors of [59] developed a ML approach for managing service function chains in MEC-enabled 5G networks. This approach combined auto-scaling prediction and optimal placement for network efficiency. The system used both classification and regression models to predict the number of user plane function instances required to meet user demands while optimizing resource usage. These SL models, particularly the regression model, achieved high accuracy in predicting auto-scaling needs by identifying patterns in network traffic data. Additionally, the system employed an integer linear programming technique to determine optimal placement of devices and service function chains. To address scalability challenges, a heuristic algorithm was also developed. Evaluations demonstrated that placing chain functions at MEC nodes based on predicted demands significantly reduced average latency within the network.

In [60], the authors designed MAPLE, a ML system for efficient VNF placement and configuration. To simplify placement decisions, MAPLE divided provider's the network infrastructure into manageable clusters. This allowed for optimized hardware and network resource allocation while meeting user

demands. The system used multi-criteria k-medoid clustering, which grouped physical resources into distinct clusters based on administrator-defined attributes. A statistical technique further refined the clustering process, reducing time and improving cluster quality. Finally, capitalizing on this network division, an ML-based placement and readjustment model dynamically mapped requested VNFs to physical resources within each cluster. This approach minimized resource waste and delivers improved QoS for users. Evaluations in a large-scale network topology showed significant improvements compared to migration techniques without clustering, including a 20% reduction in CPU usage, 25% saving in energy consumption, and 20% decrease in bandwidth utilization. From another perspective, Gupta *et al.* [55] developed the HYPER-VINES framework, a ML-based system for detecting and localizing faults that impact performance in multi-cloud VNF deployments. This framework aimed at improving the availability and reliability of VNF within cloud infrastructures. To achieve this, HYPER-VINES leveraged standard interfaces to collect performance metrics from cloud management platforms, generating large volumes of multi-source, high-dimensional data. These data underwent pre-processing to remove bias before feeding it into a two-stage detection subsystem. The first stage used a shallow ML to filter out non-faulty cases. Then, a SVM located and classified the remaining cases into imminent and already manifested faults. Evaluations using real fault logs demonstrated the framework's effectiveness and accuracy in detecting and localizing network issues that can degrade performance. Researchers in [61] proposed an architecture for assuring QoS in 5G networks using SL. The system leveraged ML to detect QoS anomalies based on historical data and network KPIs. It could also trigger automated mitigation actions and predicted future anomalies with high confidence. The system gathered 5G QoS data from devices and the RAN/core network KPIs. These data is then preprocessed, cleaned, and transformed into a unified format. A decision tree served as the core SL algorithm, building a model that correlated QoS data with KPI parameters. The anomaly detector could then identify anomalies in the network, applications, and services, consequently triggering mitigation mechanisms. Evaluations using five different traffic datasets showed the architecture achieving over 96% accuracy in anomaly detection.

The field of resource management in NGNs is vast and multifaceted, leading to the exploration of several ML approaches. RL is particularly well-suited for flexible, real-time resource orchestration. It operates in closed-loop systems, where the network continuously learns and adapts to changing user demands. RL mechanisms gather network information from monitoring agents strategically placed throughout the network architecture. On the other hand, SL finds its strengths in areas like slice admission control, VNF placement, and resource usage prediction. SL excels at complex sensing and recognition tasks due to its ability to handle the inherent complexity, size, and heterogeneity of networks from a resource management perspective. These algorithms can automatically learn patterns from data, enabling the development of sophisticated solutions. When implemented effectively, ML-based models can significantly improve performance metrics in large, distributed networks. This is because dynamic and adaptable resource management policies and decisions are crucial for optimal network function in B5G environments.

Gap analysis

The synergies between ZSM and AI models promise a future generation of B5G networks with autonomous capabilities that require minimal human intervention. While significant progress has been made during the last years, there are still multiple gaps in the design and development of fully-ZSM systems to make it possible the realization of completely autonomous network infrastructures with self-organizing and self-healing capacities.

There is an essential need to improve the data that fuel AI models within ZSM systems. The quality and accuracy of these data are of utmost importance, as inaccurate or biased data can lead the system to make suboptimal or even detrimental decisions. Better techniques for data cleansing, anomaly detection, and bias reduction will be crucial to ensure reliable predictions. Besides, techniques for interpretable AI are needed to gain insights into model behavior and build trust in ZSM decisions. When these models make decisions within ZSM systems, it can be difficult to understand the reasoning behind those

choices. This lack of explainability hinders troubleshooting, debugging, and ensuring accountability for actions taken by the system. Also, AI models trained on specific network configurations might not perform well in the heterogeneous environments found in B5G networks. Developing models that can generalize across a variety of network topologies and adapt to changing network conditions is essential for real-world ZSM deployments. Additionally, network conditions are constantly evolving. ZSM systems need to continuously learn and adapt their models to maintain optimal performance, which may be very costly in terms of CAPEX. Techniques for efficient learning and lightweight predictions are crucial to ensure models remain effective over time and able to be deployed in hardware of any kind in any point of the network continuum. In the same line, introducing AI into network management infrastructures creates new attack surfaces that can be exploited to disrupt the operation of the ZSM systems. Ensuring the security of the network and the privacy of the sensitive data they handle is vital. Robust authentication, encryption, and access control protocols are needed to safeguard against unauthorized access and manipulation.

To enable the deployment of efficient end-to-end services in NGNs infrastructures, the scalability of ZSM has to be addressed. These environments are complex and dynamic, requiring the systems to be highly modular and flexible. Distributed learning and hierarchical control architectures may help address scalability issues. Additionally, many service providers will not upgrade all their infrastructure at once. Thus, integrating ZSM solutions with legacy systems seamlessly is crucial for a smooth transition to automated network management in B5G networks. In this regards, open standards and interoperability protocols are essential to facilitate the integration. The lack of unified ML-based ZSM standards across different standardization bodies creates compatibility issues and hinders interoperability between different infrastructures. Collaboration between these bodies is needed to create a common ground. Finally, regulatory frameworks may not be fully adapted to the use of AI in network management. Therefore, establishing clear guidelines for responsible AI use, data privacy, and network security in the context of ZSM is vital.

In conclusion, while AI-based ZSM is the future for NGNs infrastructures management and orchestration, addressing these open issues is crucial to realize the dream of truly autonomous and intelligent networks. To enable the next generation of communication networks (B5G), future solutions must address both the control and data planes. Innovative packet processing pipelines, designed for efficiency and flexibility, will be essential to manage massive data volumes while ensuring low latency. These pipelines should be guided by ML-based control plane mechanisms, capable of interpreting the network's state and predicting traffic fluctuations. This predictive capability allows for proactive adjustments, preventing disruptions and maintaining optimal performance in B5G infrastructures. By ensuring data quality, security, explainability, and scalability, researchers can pave the way for a more efficient, reliable, and cost-effective network management paradigm.

3.2.2. Data plane reprogrammability

During the last years, research efforts on SDNs primarily focused on the advantages of separating the centralized control plane from the data plane. Modifications to the data plane operations has traditionally addressed by network vendors and processor designers, which limited the opportunities of researchers to actively participate in its innovation. However, the recent introduction of new technologies and advancements in operating systems, processors, network equipment, and compiler designs have significantly altered this landscape. There is a growing openness towards programmable packet handling mechanisms. This has simplified the process for researchers as they can now evaluate their solutions at line rate and in real-time, empowering them to refine their solutions before they even reach the standardization stage and bringing multiple positive implications. Firstly, it fosters a more collaborative environment where researchers and network operators can work together to develop cutting-edge data plane solutions. Secondly, by enabling pre-standardization evaluation, researchers can identify and address potential issues earlier, leading to more robust and efficient data plane implementations. Ultimately, this openness towards programmable data plane mechanisms is very promising for the future of high-performance NGNs.

To address the growing need for high-throughput networks with low latency traffic management, two promising data plane programmability technologies have emerged: P4 and eBPF. P4, an open-source domain-specific language, grants fine-grained control over how data plane-specialized devices, e.g., switches, process packets, enabling meticulous network management. eBPF, on the other hand, leverages the Linux XDP for early access to incoming packets at the operating system level. This allows for efficient and flexible traffic handling on commodity hardware, empowering even non-specialized devices to perform networking tasks effectively. By offering precise control over data plane operations and efficient traffic management on readily available hardware, respectively, P4 and eBPF pave the way for building highly efficient and programmable networks.

eBPF/XDP

The BPF technology has undergone a significant evolution within the Linux kernel since its appearance. Introduced in 1992 [62], the original version, sometimes referred to as classic BPF for clarity, is now considered obsolete. Its successor,eBPF, emerged in the Linux kernel version 3.18 and it offers a much more powerful and versatile set of functionalities. Unlike classic BPF which focused primarily on packet filtering, eBPF allows the execution of programs directly within the kernel space. This eliminates the need for modifying kernel source code or loading additional modules, enhancing its security. Furthermore, eBPF programs are written in a bytecode format, enabling safe and efficient execution at various pre-defined points within the kernel known as “hooks”. This capability grants developers the ability to inject custom code at runtime from user-space applications, essentially extending the kernel’s functionalities without compromising its core integrity. This shift from classic BPF to eBPF represents a major leap forward, transforming BPF from a simple packet filtering tool into a powerful framework for enabling advanced programmability within the Linux kernel.

Unlike traditional kernel modules that run continuously, eBPF programs operate in an event-driven manner. This efficiency stems from their association with specific hook points within the kernel as mentioned above. These hook points act as pre-defined triggers, and whenever the kernel execution reaches such a point, any eBPF programs attached to it are activated. This characteristic also makes eBPF programs stateless, handling the current situation without relying on information from previous runs. The kernel offers a rich set of predefined hooks, encompassing system calls, network events, and even kernel tracepoints, providing diverse opportunities for eBPF program intervention (see Figure 3.4). Additionally, eBPF empowers users to create custom probes. These probes function as attachment points, allowing eBPF programs to be seamlessly integrated into specific sections of kernel or user-space applications. This probe functionality unlocks a powerful monitoring capability, enabling developers to gain deep insights into the behavior of the system. In essence, eBPF’s event-driven nature and its support for custom probes combine to offer a flexible and efficient framework for program execution within the Linux kernel, facilitating diverse monitoring and optimization tasks.

eBPF offers developers multiple tools to craft programs that seamlessly integrate into the kernel. At its core, eBPF leverages a general-purpose RISC architecture. This foundational instruction set empowers developers to write eBPF programs directly in assembly language. However, for those seeking a higher-level approach, eBPF allows for programming in C language. These programs are then compiled into the eBPF bytecode using specialized compilers. Regardless of the chosen development path, all eBPF programs must undergo a rigorous verification process by the kernel’s verifier module. This verification serves as a critical safeguard for the operating system’s integrity and security. The verifier meticulously examines the program to ensure it will terminate gracefully. It also simulates the program’s execution to validate the proper usage of memory and registers throughout its operation. Additionally, the verifier scrutinizes every call made to helper functions, which are pre-defined functions within the kernel that eBPF programs can leverage for specific tasks. This verification process ensures that only safe and well-behaved programs are loaded into the kernel, preventing malicious or malfunctioning code from compromising the system’s stability. If a program fails any stage of the verification process, it will be rejected, effectively preventing it from being loaded and potentially causing harm. The whole eBPF workflow is depicted in Figure 3.5

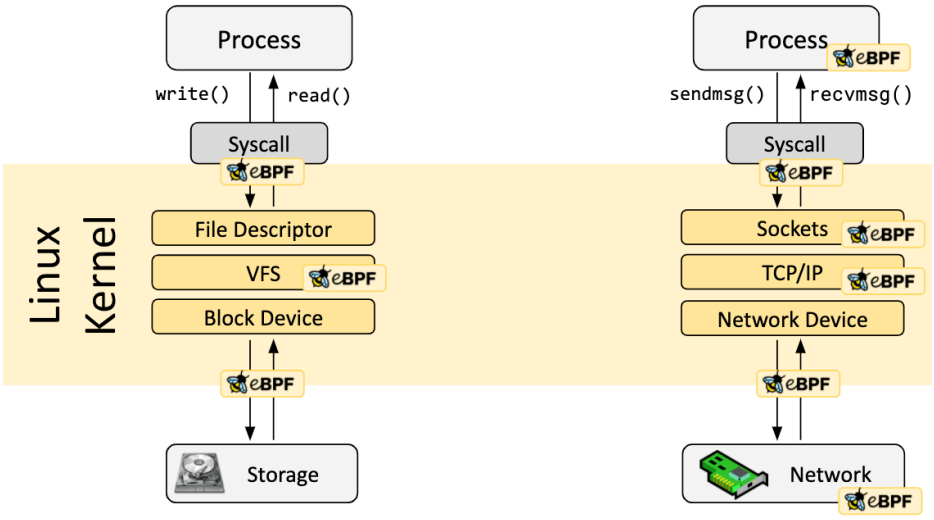

Figure 3.4: eBPF possible hooking points. Source: *ebpf.io*.

The arrival of XDP in Linux kernel version 4.8 marked a significant leap in high-performance network processing. This technology leverages the power of eBPF to enable bare-metal packet manipulation at the earliest stages of the software stack. Bypassing the typical early allocation of data structures and software queues, XDP grants eBPF programs direct access to packets received straight from the NIC through a dedicated channel. However, it is important to note that XDP is specifically designed for ingress traffic processing, acting as one of the potential hook points for eBPF program attachment. This unique position allows XDP to intervene before packets even enter the kernel's networking stack. This early access empowers XDP to perform ultra-fast processing, enabling rapid decisions when packet handling. Such decisions can involve discarding the packet entirely, modifying its content, or forwarding it for further processing within the kernel. Crucially, eBPF programs attached to the XDP hook must specify an action for each processed packet before their execution concludes. These actions offer a range of options, including dropping the packet, allowing it to proceed through the traditional kernel networking stack, or even transmitting it through the same or a different network interface. glsxdp offers flexibility in how programs are attached to the NIC through three distinct models: Generic, native, and offloaded [63]. The generic model allows the loading of eBPF/XDP programs within the Linux's standard kernel network processing path. While convenient, this approach does not unlock the full performance potential of XDP. In contrast, the native mode enables the NIC driver itself to load the program as part of its receive path. This significantly boosts performance compared to the generic model. Finally, the most advanced option involves offloading the programs directly onto the NIC hardware. However, this approach requires specialized SmartNICs with compatible firmware and drivers, making it less widely available. Therefore, the choice of model depends on the desired balance between performance and ease of implementation.

Given its unique capabilities, the eBPF/XDP synergy emerges as a promising solution for deploying network functions able to handle massive traffic volumes with strict latency requirements. This stems from XDP's ability to bypass the allocation of data structures for incoming packets at the NIC. This allows for direct manipulation of the raw packet data, enabling actions like filtering, modifying, or forwarding as discussed earlier. However, a major challenge in network management arises from potential alterations to network flows as they traverse the end-to-end path. These changes can make deployed network functions less effective in performing their intended tasks. Fortunately, the ability to compile and inject eBPF programs into the kernel at runtime presents a powerful solution to address these network architecture ossification issues. By dynamically modifying the network

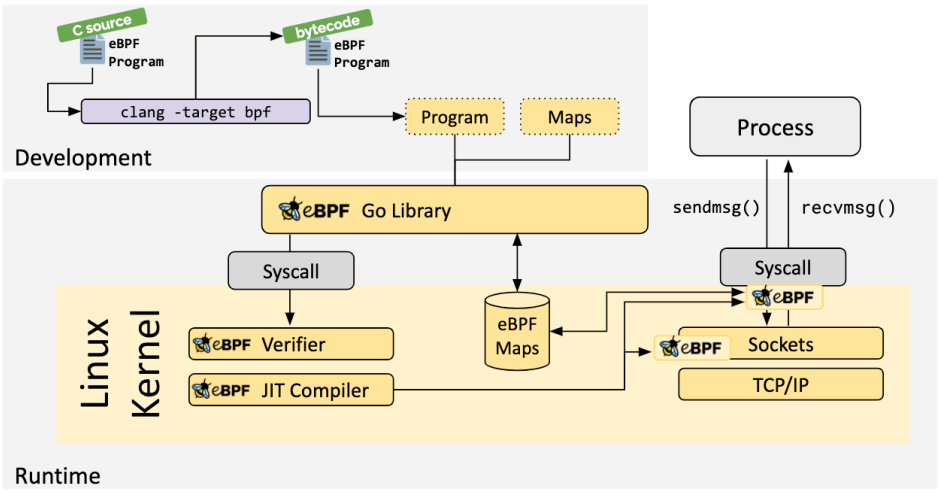

Figure 3.5: eBPF development and runtime workflow. Source: *ebpf.io*.

function code based on changing network conditions, new requirements can be accommodated without service disruptions. Even though eBPF programs are typically stateless, specific data crucial for the application can be preserved in special data structures called "maps". This allows persistence of information between program executions, making it accessible to newly injected versions of the network function. Additionally, these maps can hold application configuration details, or the configuration can be embedded within the code itself. This flexibility facilitates the creation of customized configurations that adapt to real-time network conditions and traffic patterns.

While eBPF/XDP programs offer significant advantages, their execution within the kernel space has to adhere to the strict rules set by the Linux kernel's verifier to maintain its stability and performance. A key constraint lies in the instruction limit imposed on each program. Early eBPF-enabled kernels restricted programs to a mere number of 4096 instructions, while newer versions (kernel 5.1 and above) offer more space with a 1 million instruction limit. This restriction can still pose a challenge for developing highly complex network functions. Another limitation concerns loops within eBPF programs. Only loops with a clearly defined upper bounds are permitted. This presents a barrier for network functions that rely on iterative processing, such as scanning dynamic entry lists, e.g., IP/MAC address tables, or parsing packets with variable-length headers. Fortunately, there are strategies to mitigate these limitations. For example, developers can define a maximum number of iterations at compile time using building tools. While this approach allows for iterative processing within eBPF programs, it comes at the cost of generating larger programs and sacrificing some generality as the loop's size is predetermined. These workarounds empower developers to create complex network functions within the constraints of the eBPF/XDP framework, but they can lead to larger and less versatile programs.

Beyond the previously mentioned limitations, the inherent characteristics of eBPF and XDP introduce some additional constraints. Due to the packet-driven nature of this technology, eBPF programs in the XDP hook are triggered by each incoming packet. This means eBPF/XDP network functions are not constantly running, making the management of timers or other events a non-trivial task. Additionally, handling packets before they are allocated in kernel data structures complicates the implementation of typical network protocol actions, such as temporarily holding a packet. Furthermore, the event-triggered and stateless nature limits each packet to a single action at a time. This prevents actions like forwarding a packet to multiple ports simultaneously. While this limitation can be overcome by pushing packets to deeper hooks within the kernel or even user space using maps as connection points, the additional processing introduces latency, making it unsuitable for network scenarios with strict latency requirements.

Leveraging the high-performance capabilities of eBPF/XDP programs for traffic inspection, it is

possible to deliver significant advancements over traditional Linux packet filtering mechanisms. This paves the way for efficient firewall functionalities like network access control and robust denial of service attack protection, as highlighted in [64]. Beyond security, eBPF has also found applications in user privacy. In [65], the authors proposed a mechanism that used eBPF to control DNS network traffic, ensuring user privacy with minimal overhead and without requiring application-level modifications. Overall, eBPF's versatility extended beyond performance optimization, demonstrating its potential for enhancing network security and user privacy.

The versatility of eBPF in addressing network challenges is evident in recent research. Dong *et al.* [66] proposed an eBPF program that detectedTCP network congestion in real-time. This program used eBPF probes to offer a multi-dimensional solution from the kernel's perspective, effectively addressing the prior lack of real-time network failure detection tools. The dynaprogram's dynamic probe-insertion capabilities further enhanced its usability. Another study [67] presented an eBPF-based traffic monitoring tool to measure one-way delay between machines. This tool included a novel methodology for delay measurement, with the capability to infer RTT calculation even in scenarios without time synchronization between hosts. Experiments validated the tool's microsecond-level accuracy and precision, surpassing existing solutions. Finally, work in [68] focused on improving the *upf-bpf* project, an open-source implementation of the 5G UPF using eBPF and XDP in MEC environments. This work tackled the original limitation of slow program injection time, achieving a significant 96% improvement while maintaining high performance. These studies collectively showcase eBPF's potential for real-time congestion detection, high-fidelity traffic monitoring, and flexible network function deployment.

Focusing on the efficient execution of network functions, research in [69] proposed a hardware-offloading framework that leveraged eBPF and XDP for NGNs. This framework enabled the flexible offloading of network functions directly onto SmartNICs, concretelly, Netronome Agilio SmartNICs[4]. The framework's effectiveness was validated using a QoS scenario, where tunnel identifiers from mobile traffic were used to prioritize specific users. Experiments demonstrated significant improvements in both packet processing capabilities and reduction in packet loss. These results highlighted the framework's potential for not only improving network performance but also offering the flexibility to adapt network policies on-demand.

Continuing the exploration of hardware acceleration for network functions, [70] explored a novel approach for high-performance network processing using FPGAs. Their solution, hXDP, represented a set of tools that enabled the execution of networking tasks described with eBPF programs directly on these hardware accelerators. This was achieved by enabling the Linux XDP framework to run on the FPGA, allowing for the subsequent loading and execution of eBPF programs on the NIC itself. hXDP reached this feature through a sophisticated analysis of the eBPF program's assembly code. Unnecessary instructions, irrelevant outside the Linux kernel environment, were safely removed to simplify the load the programs. Additionally, hXDP could parallelize certain parts of the code and introduce instructions specifically suited for efficient FPGA operation. Evaluations conducted using a NetFPGA NIC demonstrated the remarkable performance of hXDP. When compared to a multi-core server, hXDP achieved comparable throughput while delivering a tenfold reduction in forwarding latency. This research paved the way for significant advancements in network performance by leveraging the power of eBPF programs within the hardware acceleration capabilities of FPGAs. Another research effort by Wang *et al.* [71] introduced OXDP, a system that offloaded XDP processing to SmartNICs to achieve significant performance gains. OXDP took a specific approach, dividing the forwarding operation into two distinct planes: The data plane and the control plane. The former, responsible for the actual packet manipulation, was offloaded entirely to the SmartNIC, leveraging its hardware capabilities. The control plane, however, remained within the host machine, handling higher-level decision-making tasks. To validate OXDP's effectiveness, the researchers tested two virtual functions: A virtual switch and a virtual router. The experimental results convincingly demonstrated a substantial reduction in forwarding latency when using OXDP compared to traditional XDP implementations.

[4] https://netronome.com/

This approach highlighted the potential of offloading specific network functions to SmartNICs while maintaining control logic within the host, achieving a balance between performance and flexibility.

P4

One key concept in programmable network devices is the PISA architecture that serves as a data plane model for these devices. These elements are typically configured using P4 programs, which require compilation for the specific ASIC embedded within the target device. P4 programs and the P4 language itself are designed with the assumption that the underlying architecture consists of four key stages, depicted in Figure 3.6:

- Parsing: This stage aligns the bits within a packet with specific protocols for signaling purposes, i.e., it unpacks the information contained in the raw bit stream of the packet and interprets it based on established communication protocols.
- Match and Action: This stage allows for modifying the packet based on predefined rules. It is a decision-making and manipulation point where the packet's content can be altered based on certain criteria.
- Selection (Path or Action): Taking into account the values extracted during parsing, this stage determines the appropriate path for the packet or the specific action to be applied. The packet's journey within the network is directed based on the information extracted from the parsing stage.
- De-parsing: The final stage converts the modified bit stream back into a complete packet, ready for transmission. It is a repacking of the information after potential modifications in the match and action stage.

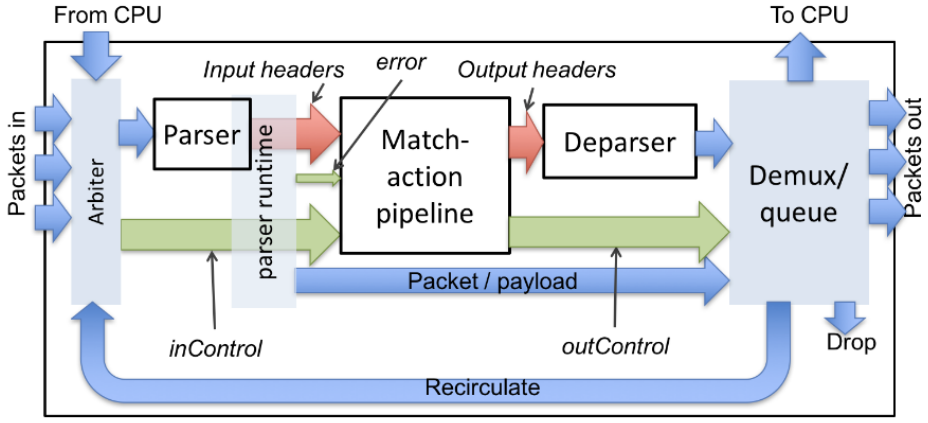

Figure 3.6: PISA reference architecture. Extracted from [72].

P4 stands as a domain-specific language specifically designed to define how packets are processed within the data plane of programmable network devices. These devices encompass a wide range, including hardware and software switches, NICs, routers, and various network appliances. The name P4 itself is inspired from the original paper that introduced the language [73]. While the initial focus of P4 was on programming network switches, its scope has significantly broadened over time. Today, it encompasses a diverse range of programmable network elements.

Traditional network switches operate with a pre-defined data plane functionality. This means that the hardware is designed to parse specific network protocols and perform specific actions based on pre-configured rules. In contrast, P4-programmable switches offer a paradigm shift. The key difference

lies in the data plane itself, where the functionality is not fixed. Instead, it is defined by the P4 program [72], which essentially instructs the switch on how to process packets, including parsing headers, performing actions based on specific criteria, and routing them accordingly. While the control plane, which manages the overall network behavior, still communicates with the data plane through similar channels as traditional switches, the underlying details have changed. The set of tables and objects that the data plane uses to make decisions are no longer fixed. Conversely, they are defined by the P4 program again. This flexibility allows for much more dynamic and customizable data planes. The P4 compiler also generates the APIs that the control plane uses to interact with the data plane. This API reflects the specific tables and objects defined in the P4 program, allowing the control plane to send instructions and receive information relevant to that specific configuration.

P4 prioritizes three key goals: (i) Reconfigurability to allow the parser and processing logic to be redefined even after deployment, (ii) protocol independence, as the hardware is agnostic to specific protocols, requiring programmers to define them alongside the parser and header processing operations, and (iii) target independence, because the underlying hardware, e.g., ASICs, FPGAs, NICs, etc. is hidden from the programmer point of view. The compiler considers the target's capabilities when transforming the P4 program into a target-specific executable. P4 programs exhibit a key performance advantage, namely, constant execution time per byte of an analyzed packet. This is true even with parser loops, as long as each iteration extracts a header field. In simpler terms, the packet size itself dictates the total processing time. This linear relationship between program complexity and total header size ensures fast packet processing across various targets, regardless of the accumulated state, e.g., number of flows or processed packets.

While P4 [73] is not the only option for programming network devices, its simplicity and the widespread availability of P4-compatible hardware and compilers have fueled its rapid adoption within the research community. This widespread adoption allows researchers to experiment and develop innovative solutions for programmable data planes with relative ease [74].

Recent research explored the potential of P4 for network diagnostics and enhanced SmartNIC functionality. In [75], the authors presented a P4-based tool for diagnosing packet loss in virtual private clouds. This tool, programmed in P4, leveraged Netronome SmartNICs to offload the packet loss detection mechanism. The experiments demonstrated that the solution can be seamlessly integrated into real-world networks without introducing significant performance overhead. Additionally, it provides valuable data for diagnosing network issues. Another noteworthy contribution came from [76], which focused on the P4-NetFPGA project. This project offered a toolchain that empowers developers to describe packet processing workflows using the P4 language, while simultaneously maintaining the option for finer-grained control through direct NetFPGA programming with languages like Verilog or VHDL. The authors proposed a custom protocol that facilitated remote population of SmartNIC routing tables and real-time retrieval of network and performance data. Validation through a TCP flow monitoring use case showcased the feasibility of this approach and highlighted how utilizing NetFPGAs as standalone P4 switches can significantly enhance platform flexibility and developer adoption. These advancements demonstrated the expanding market of P4 in programmable network devices. By leveraging SmartNICs for offloading tasks and enabling a hybrid approach combining P4 with lower-level hardware programming, researchers are laying the foundations of more efficient and adaptable network infrastructures.

Several research efforts have explored alternative methods for transporting 5G traffic, particularly focusing on mitigating the overhead associated with encapsulating GTP packets. Encapsulating GTP-U traffic is a common practice when data needs to traverse networks belonging to different operators, as it facilitates traffic handling. Traditionally, network devices perform header inspection on outer layers to steer traffic. However, this approach becomes inefficient when dealing with encapsulated packets, where the relevant information resides within the inner layers. To address this challenge, customized data planes capable of inspecting inner layer fields have been proposed. Work by Tilli *et al.* [77] investigated various alternatives to transport 5G traffic, highlighting the limitations associated with GTP-U encapsulation overhead. Another study [78] evaluated the performance of virtual evolved packet gateways using both GTP and VXLAN tunneling protocols. The evaluation was

conducted on P4-based Tofino hardware switches and software switches. Their findings revealed a direct correlation between packet size and achieved throughput, with smaller packets experiencing a significant performance drop. Additionally, the research highlighted that increased parsing complexity and per-packet actions negatively impacted the throughput and latency of software implementations, while hardware implementations remain largely unaffected. These studies underscored the limitations of traditional header inspection methods for encapsulated traffic and the potential benefits of customized data planes with deeper inspection capabilities.

Authors opf [79] proposed a double approach: A software P4 program acting as a model for control plane integration testing, and a hardware P4 program running on switches to optimize performance. This hardware program utilizes a fast path for core traffic while leveraging microservices for functionalities beyond the capabilities of high-speed switch hardware. The implementation was currently limited to research deployments on university campuses. Further studies by Paolucci *et al.* in [80] and [81] investigated the potential of P4 to unlock novel data plane functionalities not achievable with traditional SDN/NFV networks. They proposes use cases and evaluations targeting areas like advanced traffic engineering, cybersecurity, and 5G offloading. The feasibility of their approach was demonstrated through software-based performance results on platforms like BMv2 and P4rt-OVS. In [81], a P4-enhanced 5G X-haul testbed was presented, showcasing the offloading of the UPF module. This P4 implementation included functionalities like GTP encapsulation/decapsulation, configurable network steering, and real-time latency monitoring. Finally, work in [82] designed an hybrid pipeline for the UPF that leverages the strengths of both P4 targets and traditional software implementations. This approach involved running all or parts of the UPF on P4 hardware or DPDK/x86 software based on specific flow characteristics and QoS requirements. Similarly, for an hybrid gNodeB, most packet processing was handled by the P4 hardware, while functionalities not supported by P4, like Automatic Repeat Request and cryptography, were executed in DPDK/x86 software. These studies highlighted the promise of P4 for building high-performance and feature-rich data planes for B5G networks.

Gap analysis

The ever-growing demands of NGNs call for advancements in data-plane processing capabilities. Two promising solutions emerge as frontrunners: P4 and eBPF. Both offer ultra-fast packet processing and enhanced programmability for current and B5G network infrastructures. The reviewed research proposals demonstrate the potential of P4 and eBPF to improve data-plane processing within 5G infrastructures. While some P4-based solutions are already being successfully evaluated in real-world 5G deployments, showcasing the effectiveness of hardware-based traffic handling with high volumes of data, eBPF is still in its early stages. Nevertheless, eBPF offers a compelling alternative: Portable and simple data-plane programmability without the need for specialized hardware. It's important to note that the reviewed literature also explores offloading softwarized or virtualized solutions to SmartNICs or netFPGAs. This approach aims to reduce the footprint on the host CPU while enhancing throughput and reducing packet loss. However, such strategies necessitate specialized hardware, which can increase the complexity and cost of network equipment.

Even though both technologies are starting to gain popularity in the community, a crucial gap exists in the current literature. There is a lack of research comparing the performance of P4 and eBPF when integrated into real-world NGN infrastructures. Furthermore, these comparisons should also consider the influence of SDN and NFV paradigms, which are key drivers for the softwarization of modern communication architectures. Similarly, the comparison should assess the viability of integrating these technologies with ML models. This integration is key to developing intelligent network functions capable of managing the anticipated traffic demands in B5G networks. By bridging this knowledge gap, researchers will provide valuable insights into the most effective approaches for building high-performance and autonomous networks of the future.

3.2.3. Integration of AI in data plane management throughout the networking and computing continuum

Implementing AI-based technologies is essential to provide flexibility and reasoning capabilities to NGNs infrastructures, particularly through a network softwarization approach. One promising technology in this field is eBPF, which allows for efficient traffic handling on commodity hardware by enabling safe code execution within the Linux kernel as explored previously. This is especially significant for enhancing the functionality of general-purpose networking equipment, particularly in the fog-edge-cloud continuum. It is also relevant for the realization of environmentally sustainable NGNs infrastructures, as envisioned by the UN's SDGs. Enhancing resource-usage efficiency and reducing CAPEX are key drivers in designing energy-conscious B5G network architectures [83].

When combined with ML techniques, eBPF enables NGNs to conduct intelligent networking and monitoring activities across the entire infrastructure. This ability is fundamental for managing the expected high-throughput and low-latency traffic that will come with new services and applications. Furthermore, eBPF proves useful for security purposes, permitting rapid traffic analysis to detect attacks or intrusions in real-time, a critical function as new attack vectors will arise in future infrastructures and the response times will have to be almost instantaneous. Thus, eBPF emerges as a pivotal technology in B5G network infrastructures, enabling the creation of revolutionary network functions equipped with intelligent capabilities. Its inherent simplicity and flexibility allow for the deployment of these smart functions anywhere and anytime across the computing and networking continuum, a critical factor in the development of future ZSM systems within NGNs infrastructures.

Besides, IoT presents a unique and promising application area for eBPF technology, yet it also reveals its distinct challenges. IoT devices are often characterized by their limited resources, possessing minimal processing power, memory capacity, and energy reserves. These constraints create significant obstacles when attempting to implement robust and intelligent cybersecurity measures, as traditional security solutions are typically resource-intensive and may not be feasible for such constrained devices. This issue is further compounded by the emergence of a new wave of AI-based cybersecurity approaches. While these AI-driven methods show exceptional performance in detecting various types of attacks, their integration into commodity hardware remains a formidable challenge due to their computational and memory demands. This creates a critical need for innovative solutions to bridge the gap between effective cybersecurity and the resource limitations inherent to IoT devices.

In IoT deployments, every end node is a possible entry point towards the entire network infrastructure, hence improving their robustness against attacks is crucial for the overall system security. As a result, it is critical to develop defense mechanisms specifically designed for constrained hardware, taking into account their limitations and the particular threats they face in their operational environments. Since eBPF allows the efficient execution of code within the Linux kernel, it has the potential to improve the security of IoT elements by enabling the development of autonomous, sophisticated and lightweight security functions that can be fully integrated within the operative system, i.e., in its kernel. Therefore, eBPF can be used to enforce security policies at the kernel level, providing an additional layer of hardening against attacks on the end device. Thus, it is possible to build self-protection mechanisms that are tailored to the specific needs of IoT nodes, allowing them to autonomously detect and respond to security threats in a timely and efficient manner, thus reducing their dependence on the fixed infrastructure and central controllers. This will allow their direct secure integration as part of future ZSM systems in B5G networks.

In this line, as discussed before ML is considered a crucial technology to enable intelligent decision-making in network management and orchestration [84]. It allows network devices to automatically respond to changing conditions, as ML-driven mechanisms can identify anomalies, forecast network behavior, and predict failures and bottlenecks [7]. This proactive capability facilitates the automation of network management tasks, enabling real-time network optimization without human intervention. In recent years, the integration of eBPF and ML has been investigated in the literature, showcasing how their combination enables fast, proactive, and intelligent packet processing. This synergy allows network functions to automatically adapt to the demands of services and applications, significantly

enhancing network performance, flexibility, and reliability. Typically, eBPF is used to collect data from traffic flows at the Linux kernel level, while ML models operate in the user space to analyze these data and make predictions or decisions. However, this separation between traffic handling and ML processing is not optimal, as computations performed in user space generally exhibit lower performance compared to a full in-kernel implementation.

Thus, authors of [85] introduced an ML-driven framework for the dynamic selection and deployment of congestion control algorithms. Their solution leveraged two eBPF modules: One for gathering information on TCP flows and transmitting it to a user-space framework, and other one for implementing a congestion control algorithm that can be reconfigured in real-time by the framework. Experiments conducted in both simulated and real-world networks demonstrated its superior performance compared to baseline solutions. Work in [86] presented a predictive model utilizing eBPF and LSTM to monitor the status of the Linux network stack. An eBPF program was employed to track HTTP requests and responses within the kernel network stack, with these data being subsequently fed to the LSTM model to forecast future network conditions. Compared to similar approaches, this solution demonstrated higher accuracy in making real-time predictions. In [87], a kernel-level monitoring system was proposed, featuring a non-intrusive eBPF program to collect application layer traffic. The collected data were then analyzed using ML techniques to diagnose performance issues, thereby allowing for the identification of network bottlenecks. In [88], an automated Redis tuning model was created using eBPF and a random forest algorithm. The eBPF program was employed to detect various operating scenarios, and the collected data were then processed by the random forest module to rank memory parameters according to their efficiency. This optimization information was subsequently sent to the operating system to enhance hardware resource utilization. In [89], the authors introduced a solution for fingerprinting and classifying microservices. They utilized an eBPF module to trace system calls for the fingerprint process. By integrating Bayesian learning with LSTM autoencoders, their method achieved a 99% accuracy rate in fingerprinting various real-world microservices, while only incurring a 1-2% increase in CPU usage.

The integration of ML algorithms within an eBPF program is not a trivial process given the great restrictions imposed by the eBPF verifier [92]. Firstly, eBPF programs cannot make use of floating-point operations [93]. This is a great limitation given that this kind of arithmetic play a crucial role in ML algorithms. These operations, which include addition, subtraction, multiplication, and division of real numbers, are the building blocks of complex mathematical computations used in training and inference of NNs. They usually also need specialized hardware accelerators to achieve reasonable training times and inference speeds. This is why research into AI-model compression and optimization techniques, which aim to reduce the number of operations without sacrificing accuracy, is of paramount importance. In addition to hardware requirements, the energy consumption associated with them is a notable concern, particularly for applications deployed on edge devices with limited power budgets.

Moreover, the eBPF verifier imposes limitations on working with loops to ensure the termination of all programs loaded into the kernel, thus preventing potential denial of service attacks. Initially, loops were not permitted, but Linux kernel version 5.3 introduced support for bounded loops, albeit within the constraints of the maximum number of eBPF instructions allowed per program (1 million since version 5.2). Subsequently, version 5.17 added the *bpf_loop()* helper function[5], which prioritizes faster verification at the cost of some execution time. This function enables the use of larger bounded loops (up to 8 million iterations), as it is not restricted by the eBPF instruction limit. However, adhering to the verifier's stack limit remains crucial to avoid programs running excessively long due to nested loops.

Regarding the placement of eBPF programs, XDP offers flexibility in attaching them to the NIC using various models. The offloaded and native models are the most efficient, but they require the NIC driver to support this functionality [63]. However, off-the-shelf commodity hardware often features more limited NICs without such driver support. In these cases, XDP allows attaching eBPF programs

[5] https://cdn.kernel.org/pub/linux/kernel/v5.x/ChangeLog-5.17

to generic hooks that execute after the device driver. This enables advanced packet processing on constrained commodity hardware, extending the benefits of eBPF to a wider range of devices.

Gap analysis

As it was previously mentioned, eBPF is a novel technology in its infancy stages with promising capabilities to be a key enabler for autonomous managing and orchestration in NGNs infrastructures. In consequence, although its usage is beginning to extend, its presence in the literature is still limited. Nonetheless, in few years eBPF has demonstrated to be an excellent solution for monitoring and networking tasks within the Linux kernel. When combined with ML algorithms, it creates a robust tool to automatically detect and respond to network disruptions in real-time. As a main drawback, complex ML models, e.g., NNs, have not yet been integrated into the Linux kernel using eBPF capabilities, due to the difficulties of their implementation within such a restricted environment. The only preliminary work found involves implementing a simple decision tree model using a series of concatenated *if/else* statements [90]. As already discussed, the existing frameworks use eBPF programs to collect network data, which are then sent to user space where the ML algorithms process the data and make decisions [91]. This approach introduces additional overhead since the ML processing occurs outside the kernel space. Thus, research efforts should be made to overcome the technical limitations and enable the execution of ML computational tasks within the kernel space. This is crucial to enable the deployment of intelligent but lightweight network functions in commodity hardware at any point of the network, providing self-management, self-healing and self-protection in a flexible way to future communications infrastructures. Furthermore, it is also essential to address the lifecycle management of these network functions within B5G networks to assure their correct operation and optimize their contributions to ZSM systems.

3.2.4. Service-driven platforms for B5G infrastructures

Although 5G represents a significant advancement in communication networks and is often considered as the path towards a hyper-connected society, the realization of this vision remains elusive. Despite the maturity of 5G standardization efforts and the existence of early deployments, the full potential of 5G has still to be reached. This is evident in the fact that the ambitious KPIs envisioned for 5G, such as latency below 1 millisecond and bandwidth in the order of gigabits per second, are not yet widely leveraged by real-world applications. One potential reason for this discrepancy lies in the network-centric approach that has dominated the design of the 5G architecture and its associated technologies. Old releases from the 3GPP, specifically releases 15 and 16, have predominantly concentrated on improvements in the radio segment, laying the groundwork for the 5G new radio technology. While these advancements are significant, they were developed under the broad assumption of providing generic support to eMBB, uRLLC, and mMTC services. Consequently, there was a lack of clarity regarding the specific requirements of individual applications until the relatively recent release of 3GPP rel. 19, which delves into use cases and expected performance metrics. As we look towards the development of 6G, it becomes imperative to shift our focus towards user-centric services [94]. This implies to put both users and services in the center of the 6G design process from its first stages. By prioritizing the functional needs of applications and tailoring infrastructure to meet those needs, we can ensure that B5G networks are truly designed to empower the diverse range of applications and services that will shape the future of our interconnected world [95].

Undoubtedly, the design and development of B5G network infrastructures should adopt a top-down approach from the outset, considering the demands of numerous services awaiting a network infrastructure capable of supporting them. While NGNs aim to comprehensively address a wide range of vertical sectors, increased specificity in its design is crucial, achievable by focusing on individual verticals and their unique requirements. Vehicular services exemplify this need. They impose stringent QoS requirements on the underlying network infrastructure due to their inherent mobility and critical bandwidth, latency, and reliability needs [96]. The significance of the V2X vertical in the B5G

ecosystem is underscored by the attention it receives in 3GPP standards. The V2X architectural reference model, detailed in 3GPP TS 23.287 [97], is based on *PC5* (V2V) and *uu* (V2I) interfaces. This technical specification envisions application function-based service parameter provisioning for V2X communications, with the NEF serving as the entry point to the core network. Through it, B5G applications can request network operation information from user equipment, apply policies, receive location change notifications, or request traffic routing modifications, as defined in 3GPP TS 23.501 [98] and TS 23.502 [99]. This framework is crucial for ensuring compliance with QoS and security requirements within the V2X vertical. However, the specific enabling mechanisms and their design implications remain to be fully defined.

With 6G as a prominent research focus, several recent surveys have examined its developmental trajectory [6, 10, 96, 100–109]. These surveys provide extensive overviews encompassing a wide array of aspects related to the transition from 5G to 6G networks. They delve into the enabling technologies that will underpin 6G infrastructure, the diverse range of applications and services that will be made possible by these advancements, and the high-level requirements that must be met to successfully support these new capabilities. In general, these surveys tend to adopt a broad, generalist approach, striving to cover the multifaceted landscape of 6G development and deployment. A comprehensive identification of the main requirements of vehicular applications that demand cutting-edge communication capabilities is crucial. These requirements should guide the next steps in designing B5G systems. While some initial efforts have been made in previous research [6, 108–110], further refinement is necessary to establish objective and measurable metrics for future NGNs infrastructures.

NGNs designs will transcend ground-based networks, envisioning integrated architectures that interconnect space, air, land, and sea. This multi-dimensional coverage, spanning time and frequency domains, will revolutionize all modes of transportation [111]. Advancements like tactile communications, quantum computing, distributed ledger technology, THz communications, ML-based network governance and slicing, multiple radio access technologies, intelligent reflecting surfaces, and cloud-edge-fog integration will shape the future V2X ecosystem [112]. V2X encompasses communications not only among vehicles (V2V) and with infrastructure (V2I), but also with remote entities like cloud servers and even pedestrians (V2P). Consequently, existing Cellular V2X use cases [113], such as remote driving, advanced driving, vehicle platooning, extended sensors, and vehicle QoS support, will be expanded and enhanced. The potential V2X services enabled by 6G advancements are continuously expanding [10, 111, 114], promising a transformational impact on transportation and communications.

Given the rapid pace of technological innovation, it is of paramount importance to proactively identify the fundamental operational requirements that the next generation of applications in the 6G era will place on the supporting network infrastructure. These services, which will require a wide array of functionalities ranging from data transfer and real-time communication to advanced computing and artificial intelligence capabilities, will ultimately determine the performance benchmarks that the B5G infrastructure must achieve. This involves not only traditional metrics like bandwidth and latency, but also novel requirements like energy efficiency, security, and adaptability to dynamic and unpredictable environments. Therefore, understanding the specific needs and expectations of these emerging applications is crucial for guiding the design and development of 6G networks. By anticipating these requirements, we can ensure that the infrastructure is not only capable of supporting the current generation of services but it is also future-proofed to accommodate the ever-evolving landscape of technological innovation in the years to come. It is also crucial to establish specific and measurable KPIs, as they can objectively assess the performance achieved by the network infrastructure across various planes, providing valuable insights for optimization and improvement. Moreover, given this service-centric approach, which is not prevalent in current literature, it is imperative to design and implement holistic management platforms. These platforms must efficiently handle 6G-powered applications, their running environments, and associated resources (both physical and virtualized). This critical aspect has been largely overlooked in existing research.

Gap analysis

In light of this vision, the design and development of B5G networks must prioritize a holistic approach, emphasizing the creation of comprehensive infrastructures that facilitate the flexible deployment of disruptive applications across the entire fog-edge-cloud continuum. This flexibility, coupled with high network elasticity, is essential to meet the diverse and evolving demands of future applications. While 5G has made some strides in this direction from a network-centric viewpoint, the existing literature lacks platforms that fully encompass the entire B5G ecosystem, especially when considering the unique requirements of the vehicular vertical. Beyond the well-known key enabling technologies for B5G, such as Terahertz communications, MIMO systems, and ubiquitous AI, etc. [100], the development of robust and efficient orchestration platforms is essential. These platforms must be specifically designed to address the QoS needs of both users and applications across the various domains of a B5G system. This includes the ability to dynamically allocate resources, optimize network performance, and ensure seamless communication and data exchange between different components of the network. In particular, the vehicular vertical, with its stringent QoS requirements due to the critical nature of V2X communications, needs a tailored approach to ZSM systems. By incorporating intelligent algorithms and ML techniques, these platforms can proactively anticipate and respond to changes in network conditions, traffic patterns, and application demands. This will be crucial in ensuring the reliability, low latency, and high bandwidth crucial for V2X applications, such as autonomous driving, collision avoidance, and real-time traffic management.

In essence, the success of B5G infrastructures depends not only on the advancement of cutting-edge technologies but also on the development of intelligent orchestration platforms that can effectively leverage these technologies to deliver a truly transformative user experience. By adopting a holistic and service-centric approach, and by focusing on the specific needs of critical verticals like V2X, it will be possible to shape a future where B5G networks seamlessly integrate with our daily lives, empowering a wide range of applications and services that were previously unimaginable.

3.2.5. Contribution

The comprehensive review of the existing literature on the fundamental components necessary to implement ML-powered network reprogrammability solutions for ZSM in NGNs infrastructures has revealed several research gaps. These identified gaps motivate and guide the research efforts undertaken in this PhD thesis. In the following, the contributions of this thesis in each of these research areas are detailed.

Despite the substantial advancements made in recent years, significant challenges remain in the design and development of comprehensive ZSM systems. These challenges hinder the realization of fully autonomous network infrastructures with self-organizing and self-healing capabilities. There is a critical need to integrate intelligent mechanisms within NGNs infrastructures to effectively manage the diverse array of devices, services, and technologies that will converge in future hyper-connected ecosystems. In consequence, the first part of the PhD thesis analyzes and studies the state of the art in the application of techniques using ML for network orchestration and management (**Objectives 1 and 2**). This leads to the identification of the diverse network control functions essential for ZSM-managed architectures, offering insights into the most suitable ML techniques for their implementation (**R1**) [115].

In this line, B5G network infrastructures are demanding ultra-fast packet processing capabilities and enhanced programmability from the data-plane technologies (**Objective 3**). Although several alternatives are gaining attention in the community during the last years, two of the reviewed technologies demonstrated potential to be adopted within NGNs infrastructures to improve the data-plane processing capabilities: P4 and eBPF (**R2**). While both are starting to gain attention in the community, there is a lack of performance evaluations in real-world NGNs infrastructures. Thus, this thesis introduces a comprehensive performance comparison of both technologies integrated into a real-world 5G infrastructure (**R3**). This comparison examines various traffic loads and operating conditions, using a MEC architecture to process packets in a multi-tenant environment (**R4**). The

experiments demonstrate that P4 exhibits greater robustness in managing high volumes of small packet traffic. Nevertheless, the developed eBPF module achieves comparable processing latency while offering a significantly more flexible and scalable solution [116].

Despite being a relatively new technology, eBPF has shown significant promises as a key enabler for autonomous management and orchestration within NGNs. eBPF has rapidly proven to be an effective solution for monitoring and networking tasks within the Linux kernel. When combined with ML algorithms, it becomes a powerful tool for real-time, automatic detection and response to network disruptions. A notable limitation is that complex ML models, such as NNs, have not yet been successfully integrated into the Linux kernel using eBPF due to the challenges of implementing them in this constrained environment. Therefore, research should prioritize overcoming the technical limitations that prevent the execution of ML computational tasks within the kernel space (**Objective 4**). This advancement is crucial for enabling the deployment of intelligent yet lightweight network functions on standard hardware throughout the network (**Objective 5**). In consequence, in this PhD thesis, a solution that seamlessly integrates intelligent traffic inspection models directly into the Linux kernel is developed. By leveraging the capabilities of eBPF, rapid packet processing is combined with ML-based intelligent decision-making within the same level (**R5**). The approach is validated in commodity hardware and it demonstrates its usefulness to save resources on the device performing the computation and, also, it significantly reduces processing latencies, which is crucial in edge-enabled B5G systems (**R6**) [117].

Finally, the success of B5G infrastructures hinges not only on the advancement of cutting-edge technologies but also on the design of network architectures that can effectively integrate these technologies to provide an innovative user experience (**Objective 6**). While 6G aims to encompass a wide array of vertical sectors, achieving a truly effective design needs a more targeted approach, focusing on specific verticals and their unique requirements (**Objective 7**). For this reason, this PhD thesis explores these needs from the point of view of the vehicular vertical. The insights gained from the aforementioned research lines are synthesized to envision the pivotal networking and computing elements that will direct the design of future 6G infrastructures (**R7**). The demanding and complex automotive sector, with its stringent requirements, is chosen as a focal point. It is explored how the future applications and services envisioned for this sector will significantly influence the definition and implementation of 6G networks. Furthermore, drawing upon the conducted analysis, a conceptual design for a ZSM-based 6G platform is presented (**R8**). This platform is envisioned to comprehensively manage the entire lifecycle of multi-domain V2X services and applications, encompassing their deployment, monitoring, optimization, and evolution [118].

3.3. Lessons Learned and Conclusions

While 5G is currently rolling out globally, the concept of NGNs has already begun to capture the attention of researchers, industry leaders, and the general public. Discussions and investigations into B5G networks are in full swing, encompassing a wide range of topics, from their conceptual framework and underlying architecture to the enabling technologies that will power these NGN infrastructures. There are numerous compelling reasons to justify the pursuit of a new generational jump in communication networks as we understand them. First and foremost, we are witnessing an unprecedented acceleration in the development and adoption of technologies that demand immense resources. Autonomous driving, extended reality, and precision medicine are just a few examples of fields that are rapidly advancing and requiring increasingly sophisticated communication networks to function effectively. Simultaneously, the trend towards miniaturization of smart devices continues relentless. These combined forces will inevitably strain the computing and communication capabilities of 5G networks, making it clear that a more powerful and adaptable successor is necessary to meet future demands.

Despite the advancements brought about by 5G, it still falls short in certain crucial areas. Providing autonomous and high-performance network infrastructures remains a challenge for 5G. This is a

significant limitation, as the vision of NGNs is to create a completely self-managed and self-orchestrated network based on the ZSM paradigm, where the self-governance becomes a reality, reducing the human intervention to near zero and providing network capabilities never seen before. Another key factor in NGNs lies in the driving forces behind their development. While the evolution of mobile networks from 1G to 5G has been largely propelled by societal and commercial factors, the trajectory of NGNs is expected to be shaped by a broader range of influences. Political, economic, social, and even military considerations are likely to play a significant role in the development of NGNs, reflecting the growing recognition of its potential to transform various aspects of our lives and society as a whole.

Although AI-powered ZSM holds immense promise for the management and orchestration of NGNs, resolving several open challenges is paramount to achieving truly autonomous and intelligent networks. To usher in the era of B5G communication networks, future solutions must encompass both the control and data planes. Innovative packet processing pipelines, engineered for efficiency and adaptability, will be indispensable for managing the enormous data volumes anticipated while maintaining minimal latency. These pipelines should be steered by ML-based control plane mechanisms that can infer the network's future state and anticipate traffic fluctuations. This predictive capability enables proactive adjustments, preventing disruptions and upholding optimal performance within B5G infrastructures. By focusing on data quality, security, explainability, and scalability, next research efforts should pave the way for a network management paradigm that is not only more efficient and reliable but also cost-effective.

Besides, the escalating demands of NGNs need advancements in data-plane processing capabilities. P4 and eBPF have emerged as leading contenders, both promising ultra-fast packet processing and enhanced programmability for current and B5G network infrastructures. Despite their growing popularity, there is a significant gap in the current research landscape. There is a lack of studies comparing the performance of P4 and eBPF when integrated into real-world NGN infrastructures. Critically, these comparisons should also account for the influence of SDN and NFV paradigms, which are central to the softwarization of modern communication architectures. Additionally, the assessment should evaluate the feasibility of integrating these technologies with ML models. This integration is essential for the development of intelligent network functions capable of handling the massive traffic volumes expected in NGNs. By addressing this knowledge gap, invaluable insights into the most effective strategies can be offered for constructing high-performance and autonomous networks of the future.

In this line, eBPF, a relatively new technology in its early stages, has shown significant promise as a key enabler for autonomous management and orchestration within B5G. While its use is expanding, its presence in research literature remains limited. However, even in its infancy, eBPF has proven to be an excellent solution for monitoring and networking within the Linux kernel. When combined with ML algorithms, it forms a powerful tool for automatically detecting and responding to network disruptions in real time. A primary challenge is that complex ML models have not yet been successfully integrated into the Linux kernel using eBPF capabilities. This is due to the difficulties of implementing such models within a restricted environment. To unlock the full potential of eBPF, research efforts should focus on overcoming these technical limitations and enabling the execution of ML computational tasks within the kernel space. This is crucial for deploying intelligent yet lightweight network functions on commodity hardware at any point in the network, thereby providing self-management, self-healing, and self-protection in a flexible manner for future communications infrastructures.

Furthermore, the design and development of B5G networks must prioritize a holistic approach, focusing on building comprehensive infrastructures that enable the flexible deployment of the developed smart network functions across the entire fog-edge-cloud continuum. This adaptability, combined with high network elasticity, is crucial for meeting the diverse and ever-changing requirements of future applications. Ultimately, the success of NGNs infrastructures hinges not only on the advancement of cutting-edge technologies but also on the development of intelligent orchestration platforms. These platforms must effectively leverage these technologies to deliver a truly transformative and seamless user experience.

Finally, NGNs are not merely an incremental upgrade over 5G, but rather a paradigm shift in

communication technology. It transcends the traditional focus on increasing data rates and other basic performance metrics, aiming instead to achieve real-time information sharing among all entities, considering both humans and machines. This includes bridging the digital divide, ensuring that everyone has access to the benefits of this transformative technology. Experts envision NGNs as the driving force behind information interaction and social life beyond 2030, leading to the creation of user-centric services and network infrastructures characterized by ML-driven intelligence, high dynamism, extreme heterogeneity, and the seamless integration of all things into the computing and networking continuum.

CHAPTER 4

Publications Composing the PhD Thesis

4.1. Machine learning-based zero-touch network and service management: a survey

Title	Machine learning-based zero-touch network and service management: a survey
Authors	Jorge Gallego-Madrid and Ramon Sanchez-Iborra and Pedro M. Ruiz and Antonio F. Skarmeta
Type	Journal
Journal	Digital Communications and Networks
Impact Factor	6.3 - Q1 (2021)
Publisher	KeAi Publishing LTD
Pages	105-123
Volume	8
Issue	2
Year	2021
Month	September
ISSN	2468-5925
DOI	https://doi.org/10.1016/j.dcan.2021.09.001
URL	https://www.sciencedirect.com/science/article/pii/S2352864821000614
State	Published
Author's contribution	The PhD student, Jorge Gallego-Madrid, is the main author of the paper

Journal details: Digital Communications and Networks
ISSN: 2468-5925
Publisher: KeAi Publishing LTD
Impact factor (2021): 6.3
Website: `https://www.sciencedirect.com/journal/digital-communications-and-networks`

Authors – Personal details	
Name	**Jorge Gallego-Madrid**
Position	PhD student at Department of Information and Communications Engineering
University	University of Murcia
Name	**Dr. Ramon Sanchez-Iborra**
Position	Associate Professor at University Center of Defense at the Spanish Air Force Academy
University	University Center of Defense at the Spanish Air Force Academy
Name	**Dr. Pedro M. Ruiz**
Position	Associate Professor at Department of Information and Communications Engineering
University	University of Murcia
Name	**Dr. Antonio F. Skarmeta Gómez**
Position	Full Professor at Department of Information and Communications Engineering
University	University of Murcia

Abstract

The exponential growth of mobile applications and services during the last years has challenged the existing network infrastructures. Consequently, the arrival of multiple management solutions to cope with this explosion along the end-to-end network chain has increased the complexity in the coordinated orchestration of different segments composing the whole infrastructure. The Zero-touch Network and Service Management (ZSM) concept has recently emerged to automatically orchestrate and manage network resources while assuring the Quality of Experience (QoE) demanded by users. Machine Learning (ML) is one of the key enabling technologies that many ZSM frameworks are adopting to bring intelligent decision making to the network management system. This paper presents a comprehensive survey of the state-of-the-art application of ML-based techniques to improve ZSM performance. To this end, the main related standardization activities and the aligned international projects and research efforts are deeply examined. From this dissection, the skyrocketing growth of the ZSM paradigm can be observed. Concretely, different standardization bodies have already designed reference architectures to set the foundations of novel automatic network management functions and resource orchestration. Aligned with these advances, diverse ML techniques are being currently exploited to build further ZSM developments in different aspects, including multi-tenancy management, traffic monitoring, and architecture coordination, among others. However, different challenges, such as the complexity, scalability, and security of ML mechanisms, are also identified, and future research guidelines are provided to accomplish a firm development of the ZSM ecosystem.

4.2. Fast traffic processing in multi-tenant 5G environments: A comparative performance evaluation of P4 and eBPF technologies

Title	Fast traffic processing in multi-tenant 5G environments: A comparative performance evaluation of P4 and eBPF technologies
Authors	Jorge Gallego-Madrid and Alejandro Molina-Zarca and Ramon Sanchez-Iborra and Jordi Ortiz and Antonio F. Skarmeta
Type	Journal
Journal	Engineering Science and Technology, an International Journal
Impact Factor	5.7 - Q1 (2022)
Publisher	Elsevier
Pages	101678
Volume	52
Issue	-
Year	2024
Month	March
ISSN	2215-0986
DOI	https://doi.org/10.1016/j.jestch.2024.101678
URL	https://www.sciencedirect.com/science/article/pii/S2215098624000648
State	Published
Author's contribution	The PhD student, Jorge Gallego-Madrid, is the main author of the paper

Journal details: Engineering Science and Technology, an International Journal
ISSN: 2215-0986
Publisher: Elsevier
Impact factor (2022): 5.7
Website: `https://www.sciencedirect.com/journal/engineering-science-and-technology-an-international-journal`

Authors – Personal details	
Name	**Jorge Gallego-Madrid**
Position	PhD student at Department of Information and Communications Engineering
University	University of Murcia
Name	**Dr. Alejandro Molina-Zarca**
Position	Assistant Professor at University Center of Defense at the Spanish Air Force Academy
University	University Center of Defense at the Spanish Air Force Academy
Name	**Dr. Ramon Sanchez-Iborra**
Position	Associate Professor at Department of Information and Communications Engineering
University	University of Murcia
Name	**Dr. Jordi Ortiz**
Position	Assistant Professor at University Center of Defense at the Spanish Air Force Academy
University	University Center of Defense at the Spanish Air Force Academy
Name	**Dr. Antonio F. Skarmeta Gómez**
Position	Full Professor at Department of Information and Communications Engineering
University	University of Murcia

Abstract

Although the softwarization of network infrastructures through the use of Software Defined Networking (SDN) and Network Function Virtualization (NFV) has set the foundations of future communication architectures, the efficient handling of high throughput traffic while maintaining latency requirements still remains a challenge. In this work, we explore two arising technologies that aim at reducing networking tasks' latency while dealing with high levels of traffic volume, namely, Programming Protocol-independent Packet Processors (P4) and the extended Berkeley Packet Filter (eBPF). We present a review of the latest advances in the use of both technologies and we provide a discussion on their advantages and disadvantages. As the main contribution of the paper, we showcase an extensive performance evaluation of these technologies under different traffic conditions. To do so, we implement a fast traffic processing network function operating in a real 5G Stand Alone (SA) network. Obtained results confirm, as expected, the high performance attained using dedicated hardware programmed by P4, in contrast to eBPF-based solution's poorer results while handling similar throughputs. Nevertheless, eBPF allows similar packet-processing times than P4, therefore qualifying it as a perfectly scalable solution on commodity hardware even as a virtual function, which paves the way for the realization of autonomous, flexible and cost-effective next-generation network infrastructures.

4.3. Machine learning-powered traffic processing in commodity hardware with eBPF

Title	Machine learning-powered traffic processing in commodity hardware with eBPF
Authors	Jorge Gallego-Madrid and Irene Bru-Santa and Alvaro Ruiz-Rodenas and Ramon Sanchez-Iborra and Antonio F. Skarmeta
Type	Journal
Journal	Computer Networks
Impact Factor	5.6 - Q1 (2022)
Publisher	Elsevier
Pages	110295
Volume	243
Issue	-
Year	2024
Month	April
ISSN	1389-1286
DOI	https://doi.org/10.1016/j.comnet.2024.110295
URL	https://www.sciencedirect.com/science/article/pii/S1389128624001270
State	Published
Author's contribution	The PhD student, Jorge Gallego-Madrid, is the main author of the paper

Journal details: Computer Networks
ISSN: 1389-1286
Publisher: Elsevier
Impact factor (2022): 5.6
Website: `https://www.sciencedirect.com/journal/computer-networks`

Authors – Personal details	
Name	**Jorge Gallego-Madrid**
Position	PhD student at Department of Information and Communications Engineering
University	University of Murcia
Name	**Irene Bru-Santa**
Position	Student at Department of Information and Communications Engineering
University	University of Murcia
Name	**Alvaro Ruiz-Rodenas**
Position	Student at Department of Information and Communications Engineering
University	University of Murcia
Name	**Dr. Ramon Sanchez-Iborra**
Position	Associate Professor at Department of Information and Communications Engineering
University	University of Murcia
Name	**Dr. Antonio F. Skarmeta Gómez**
Position	Full Professor at Department of Information and Communications Engineering
University	University of Murcia

Abstract

Network softwarization is paving the way for the design and development of Next-Generation Networks (NGNs), which are demanding profound improvements to existing communication infrastructures. Two of the fundamental pillars of NGNs are flexibility and intelligence to create elastic network functions capable of managing complex communication systems in an efficient and cost-effective way. In this sense, the extended Berkeley Packet Filter (eBPF) is a state-of-the-art solution that enables low-latency traffic processing within the Linux kernel in commodity hardware. When combined with Machine Learning (ML) algorithms, it becomes a promising enabler to perform smart monitoring and networking tasks at any required place of the fog-edge-cloud continuum. In this work, we present a solution that leverages eBPF to integrate ML-based intelligence with fast packet processing within the Linux kernel, enabling the execution of complex computational tasks in a flexible way, saving resources and reducing processing latencies. A real implementation and a series of experiments have been carried out in an Internet of Things (IoT) scenario to evaluate the performance of the solution to detect attacks in a 6LowPAN system. The performance of the in-kernel implementation shows a considerable reduction in the execution time (-97%) and CPU usage (-6%) of a Multi-Layer Perceptron (MLP) model in comparison with a user space development approach; thus positioning our proposal as a promising solution to embed ML-powered fast packet processing within the Linux kernel.

4.4. The role of vehicular applications in the design of future 6G infrastructures

Title	The role of vehicular applications in the design of future 6G infrastructures
Authors	Jorge Gallego-Madrid and Ramon Sanchez-Iborra and Jordi Ortiz and Jose Santa
Type	Journal
Journal	ICT Express
Impact Factor	5.4 - Q2 (2022)
Publisher	Elsevier
Pages	556-570
Volume	9
Issue	4
Year	2023
Month	August
ISSN	2405-9595
DOI	https://doi.org/10.1016/j.icte.2023.03.011
URL	https://www.sciencedirect.com/science/article/pii/S2405959523000383
State	Published
Author's contribution	The PhD student, Jorge Gallego-Madrid, is the main author of the paper

Journal details: ICT Express
ISSN: 2405-9595
Publisher: Elsevier
Impact factor (2022): 5.4
Website: `https://www.sciencedirect.com/journal/ict-express`

Authors – Personal details	
Name	**Jorge Gallego-Madrid**
Position	PhD student at Department of Information and Communications Engineering
University	University of Murcia
Name	**Dr. Ramon Sanchez-Iborra**
Position	Associate Professor at Department of Information and Communications Engineering
University	University of Murcia
Name	**Dr. Jordi Ortiz**
Position	Assistant Professor at University Center of Defense at the Spanish Air Force Academy
University	University Center of Defense at the Spanish Air Force Academy
Name	**Dr. Jose Santa**
Position	Associate Professor at Technical University of Cartagena
University	Technical University of Cartagena

Abstract

A great lack of 5G design is the traditional bottom-up development of network evolution, which has not effectively considered the requirements of applications and, particularly, vehicle to everything (V2X) applications. This paper provides a service-centric approach towards 6G V2X, with a concise overview of the upcoming hyper-connected vehicular ecosystem and its integration in the whole 6G fabric, analysing its particular infrastructure needs, as a way to reach key performance indicators (KPIs). We also present a 6G-oriented platform design able to manage the life-cycle of V2X applications across different domains by means of intelligent orchestration decisions.

Chapter 5

Bibliography

5.1. References

[1] V. Sciancalepore, F. Z. Yousaf, and X. Costa-Perez, "Z-TORCH: An Automated NFV Orchestration and Monitoring Solution," *IEEE Transactions on Network and Service Management*, vol. 15, no. 4, pp. 1292–1306, dec 2018.

[2] O. U. Akgul, I. Malanchini, and A. Capone, "Dynamic Resource Trading in Sliced Mobile Networks," *IEEE Transactions on Network and Service Management*, vol. 16, no. 1, pp. 220–233, mar 2019.

[3] R. Ali, Y. B. Zikria, A. K. Bashir, S. Garg, and H. S. Kim, "URLLC for 5G and Beyond: Requirements, Enabling Incumbent Technologies and Network Intelligence," *IEEE Access*, vol. 9, pp. 67 064–67 095, 2021.

[4] "European Vision for the 6G Network Ecosystem < 5G-PPP." [Online]. Available: https://5g-ppp.eu/european-vision-for-the-6g-network-ecosystem/

[5] M. Noor-A-Rahim, Z. Liu, H. Lee, M. O. Khyam, J. He, D. Pesch, K. Moessner, W. Saad, and H. V. Poor, "6G for Vehicle-to-Everything (V2X) Communications: Enabling Technologies, Challenges, and Opportunities," *Proceedings of the IEEE*, vol. 110, no. 6, pp. 712–734, jun 2022.

[6] Y. Lu and X. Zheng, "6G: A survey on technologies, scenarios, challenges, and the related issues," *Journal of Industrial Information Integration*, vol. 19, p. 100158, sep 2020.

[7] A. Shahraki, T. Ohlenforst, and F. Kreyß, "When machine learning meets Network Management and Orchestration in Edge-based networking paradigms," *Journal of Network and Computer Applications*, vol. 212, p. 103558, mar 2023.

[8] G. Montenegro, J. Hui, D. Culler, and N. Kushalnagar, "Rfc ft-ietf-6lowpan-format: Transmission of ipv6 packets over ieee 802.15.4 networks," Sep 2007. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc4944

[9] J. Nordby, M. Cooke, and A. Horvath, "emlearn: Machine Learning inference engine for Microcontrollers and Embedded Devices," Mar. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2589394

[10] G. Kirubasri, S. Sankar, D. Pandey, B. K. Pandey, H. Singh, and R. Anand, "A Recent Survey on 6G Vehicular Technology, Applications and Challenges," in *9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. IEEE, sep 2021, pp. 1–5.

[11] H. Guo, X. Zhou, J. Liu, and Y. Zhang, "Vehicular intelligence in 6G: Networking, communications, and computing," *Vehicular Communications*, vol. 33, p. 100399, jan 2022.

[12] Huawei, "6G: The Next Horizon," 2021. [Online]. Available: https://www.huawei.com/en/huaweitech/future-technologies/6g-white-paper

[13] Ericsson, "With 6G, let's connect a cyber-physical world - Ericsson," 2022. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/white-papers/a-research-outlook-towards-6g

[14] M. Liyanage, Q.-V. Pham, K. Dev, S. Bhattacharya, P. Kumar, R. Maddikunta, T. Reddy Gadekallu, and G. Yenduri, "A survey on Zero touch network and Service Management (ZSM) for 5G and beyond networks," *Journal of Network and Computer Applications*, vol. 203, p. 103362, 2022. [Online]. Available: http://creativecommons.org/licenses/by/4.0/

[15] C. Benzaid and T. Taleb, "AI-Driven Zero Touch Network and Service Management in 5G and Beyond: Challenges and Research Directions," *IEEE Network*, vol. 34, no. 2, pp. 186–194, mar 2020.

[16] G. Chollon, D. Ayed, R. A. Garriga, A. M. Zarca, A. Skarmeta, M. Christopoulou, W. Soussi, G. Gur, and U. Herzog, "ETSI ZSM Driven Security Management in Future Networks," *Proceedings - 2022 IEEE Future Networks World Forum, FNWF 2022*, pp. 334–339, 2022.

[17] E. Coronado, R. Behravesh, T. Subramanya, A. Fernandez-Fernandez, M. S. Siddiqui, X. Costa-Perez, and R. Riggio, "Zero Touch Management: A Survey of Network Automation Solutions for 5G and 6G Networks," *IEEE Communications Surveys and Tutorials*, vol. 24, no. 4, pp. 2535–2578, 2022.

[18] ETSI, "GS ZSM 002 - V1.1.1 - Zero-touch network and Service Management (ZSM); Reference Architecture," 2019. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

[19] E. I. ENI, "GS ENI 005 - V3.1.1 - Experiential Networked Intelligence (ENI); System Architecture," 2023. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

[20] E. I. ENI, "GR ENI 004 - V3.1.1 - Experiential Networked Intelligence (ENI); Terminology," 2023. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

[21] E. I. ENI, "GS ENI 002 - V3.2.1 - Experiential Networked Intelligence (ENI); ENI requirements," 2023. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

[22] E. I. ENI, "GS ENI 001 - V3.2.1 - Experiential Networked Intelligence (ENI); ENI Use Cases," 2023. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

[23] E. I. ENI, "GR ENI 009 - V1.2.1 - Experiential Networked Intelligence (ENI); Definition of data processing mechanisms," 2023. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

[24] ETSI, "TS 103 195-2 - V1.1.1 - Autonomic network engineering for the self-managing Future Internet (AFI); Generic Autonomic Network Architecture; Part 2: An Architectural Reference Model for Autonomic Networking, Cognitive Networking and Self-Management," 2018. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

[25] ITU-T, "Y.3172 : Architectural framework for machine learning in future networks including IMT-2020." [Online]. Available: https://www.itu.int/rec/T-REC-Y.3172-201906-I/en

[26] ITU-T, "Y.Sup55 : ITU-T Y.3170-series - Machine learning in future networks including IMT-2020: use cases." [Online]. Available: https://www.itu.int/rec/T-REC-Y.Sup55-201910-I

[27] ITU-T, "Y.3173 : Framework for evaluating intelligence levels of future networks including IMT-2020." [Online]. Available: https://www.itu.int/rec/T-REC-Y.3173-202002-I

[28] ITU-T, "Y.3174 : Framework for data handling to enable machine learning in future networks including IMT-2020." [Online]. Available: https://www.itu.int/rec/T-REC-Y.3174-202002-I

[29] ITU-T, "Y.3176 : Machine learning marketplace integration in future networks including IMT-2020." [Online]. Available: https://www.itu.int/rec/T-REC-Y.3176-202009-P

[30] TSGC, "TS 129 552 - V17.3.0 - 5G; 5G System; Network Data Analytics signalling flows; Stage 3 (3GPP TS 29.552 version 17.3.0 Release 17)," 2023. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

[31] TSGS, "Universal Mobile Telecommunications System (UMTS); LTE; Telecommunication management; Self-Organizing Networks (SON); Concepts and requirements (3GPP TS 32.500 version 8.0.0 Release 8)," 2009. [Online]. Available: http://portal.etsi.org/chaircor/ETSI_support.asp

[32] TSGS, "TS 128 313 - V16.0.0 - 5G; Self-Organizing Networks (SON) for 5G networks (3GPP TS 28.313 version 16.0.0 Release 16)," 2020. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

[33] L. V. Le, B. S. P. Lin, L. P. Tung, and D. Sinh, "SDN/NFV, Machine Learning, and Big Data Driven Network Slicing for 5G," *IEEE 5G World Forum, 5GWF 2018 - Conference Proceedings*, pp. 20–25, 2018.

[34] Lim, Kim, Kim, Hong, and Han, "Payload-Based Traffic Classification Using Multi-Layer LSTM in Software Defined Networks," *Applied Sciences*, vol. 9, no. 12, p. 2550, jun 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/12/2550

[35] T. V. Phan, S. T. Islam, T. G. Nguyen, and T. Bauschert, "Q-DATA: Enhanced Traffic Flow Monitoring in Software-Defined Networks applying Q-learning," in *2019 15th International Conference on Network and Service Management (CNSM)*. IEEE, oct 2019, pp. 1–9. [Online]. Available: http://arxiv.org/abs/1909.01544https://ieeexplore.ieee.org/document/9012727/

[36] D. Szostak and K. Walkowiak, "Machine learning methods for traffic prediction in dynamic optical networks with service chains," in *International Conference on Transparent Optical Networks*, vol. 2019-July. IEEE Computer Society, jul 2019.

[37] X. Chen, B. Li, R. Proietti, Z. Zhu, and S. J. Ben Yoo, "Multi-agent deep reinforcement learning in cognitive inter-domain networking with multi-broker orchestration," in *Optics InfoBase Conference Papers*, vol. Part F160-, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8696667

[38] G. Liu, K. Zhang, X. Chen, H. Lu, J. Guo, J. Yin, R. Proietti, Z. Zhu, and S. J. B. Yoo, "Hierarchical Learning for Cognitive End-to-End Service Provisioning in Multi-Domain Autonomous Optical Networks," *Journal of Lightwave Technology*, vol. 37, no. 1, pp. 218–225, jan 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8550664/

[39] R. Proietti, X. Chen, K. Zhang, G. Liu, M. Shamsabardeh, A. Castro, L. Velasco, Z. Zhu, and S. J. Ben Yoo, "Experimental demonstration of machine-learning-aided qot estimation in multi-domain elastic optical networks with alien wavelengths," *Journal of Optical Communications and Networking*, vol. 11, no. 1, pp. A1–A10, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8651192

[40] M. Shin and M. Y. Chung, "Learning-based Distributed Multi-channel Dynamic Access for Cellular Spectrum Sharing of Multiple Operators," in *2019 25th Asia-Pacific Conference on Communications (APCC)*. IEEE, nov 2019, pp. 384–387. [Online]. Available: https://ieeexplore.ieee.org/document/9026533/

[41] K. Thar, T. Z. Oo, Y. K. Tun, D. H. Kim, K. T. Kim, and C. S. Hong, "A Deep Learning Model Generation Framework for Virtualized Multi-Access Edge Cache Management," *IEEE Access*, vol. 7, pp. 62 734–62 749, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8712457/

[42] Z. Zhang, L. Ma, K. Poularakis, K. K. Leung, J. Tucker, and A. Swami, "MACS: Deep Reinforcement Learning based SDN Controller Synchronization Policy Design," *2019 IEEE 27th International Conference on Network Protocols (ICNP)*, vol. 2019-Octob, pp. 1–11, sep 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8888034/http://arxiv.org/abs/1909.09063

[43] H. D. R. Albonda and J. Pérez-Romero, "Reinforcement Learning-Based Radio Access Network Slicing for a 5G System with Support for Cellular V2X," in *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, vol. 291. Springer Verlag, jun 2019, pp. 262–276. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-25748-4_20

[44] C.-J. Chun, J.-M. Kang, and I.-M. Kim, "Deep Learning-Based Joint Pilot Design and Channel Estimation for Multiuser MIMO Channels," *IEEE Communications Letters*, vol. 23, no. 11, pp. 1999–2003, nov 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8813060/

[45] M. Guan, Z. Wu, Y. Cui, X. Cao, L. Wang, J. Ye, and B. Peng, "An intelligent wireless channel allocation in HAPS 5G communication system based on reinforcement learning," *Eurasip Journal on Wireless Communications and Networking*, vol. 2019, no. 1, pp. 1–9, dec 2019. [Online]. Available: https://link.springer.com/articles/10.1186/s13638-019-1463-8https://link.springer.com/article/10.1186/s13638-019-1463-8

[46] J. M. Kang, C. J. Chun, and I. M. Kim, "Deep Learning Based Channel Estimation for MIMO Systems with Received SNR Feedback," *IEEE Access*, vol. 8, pp. 121 162–121 181, 2020.

[47] D. Lynch, T. Saber, S. Kucera, H. Claussen, and M. O'Neill, "Evolutionary learning of link allocation algorithms for 5G heterogeneous wireless communications networks," in *GECCO 2019 - Proceedings of the 2019 Genetic and Evolutionary Computation Conference*. New York, NY, USA: Association for Computing Machinery, Inc, jul 2019, pp. 1258–1265. [Online]. Available: https://dl.acm.org/doi/10.1145/3321707.3321853

[48] J. Ma, S. Hasegawa, S.-J. Kim, and M. Hasegawa, "A Reinforcement-Learning-Based Distributed Resource Selection Algorithm for Massive IoT," *Applied Sciences*, vol. 9, no. 18, p. 3730, sep 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/18/3730

[49] R. Politanskyi and M. Klymash, "Application of Artificial Intelligence in Cognitive Radio for Planning Distribution of Frequency Channels," in *2019 3rd International Conference on Advanced Information and Communications Technologies, AICT 2019 - Proceedings*. Institute of Electrical and Electronics Engineers Inc., jul 2019, pp. 390–394.

[50] M. R. Raza, C. Natalino, P. Ohlen, L. Wosinska, and P. Monti, "Reinforcement Learning for Slicing in a 5G Flexible RAN," *Journal of Lightwave Technology*, vol. 37, no. 20, pp. 5161–5169, 2019.

[51] R. M. Sandoval, S. Canovas-Carrasco, A. J. Garcia-Sanchez, and J. Garcia-Haro, "Smart usage of multiple rat in IoT-oriented 5G networks: A reinforcement learning approach," *10th ITU Academic Conference Kaleidoscope: Machine Learning for a 5G Future, ITU K 2018*, pp. 1–8, 2018.

[52] R. Sanchez-Iborra and A. F. Skarmeta, "TinyML-Enabled Frugal Smart Objects: Challenges and Opportunities," *IEEE Circuits and Systems Magazine*, vol. 20, no. 3, pp. 4–18, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9166461/

[53] M. Bunyakitanon, X. Vasilakos, R. Nejabati, and D. Simeonidou, "End-to-End Performance-based Autonomous VNF Placement with adopted Reinforcement Learning," *IEEE Transactions on Cognitive Communications and Networking*, pp. 1–1, apr 2020.

[54] F. De Vita, D. Bruneo, A. Puliafito, G. Nardini, A. Virdis, and G. Stea, "A deep reinforcement learning approach for data migration in multi-access edge computing," *10th ITU Academic Conference Kaleidoscope: Machine Learning for a 5G Future, ITU K 2018*, pp. 1–8, 2018.

[55] L. Gupta, T. Salman, R. Das, A. Erbad, R. Jain, and M. Samaka, "HYPER-VINES: A HYbrid Learning Fault and Performance Issues ERadicator for Virtual NEtwork Services over Multi-Cloud Systems," in *2019 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, feb 2019, pp. 141–147. [Online]. Available: https://ieeexplore.ieee.org/document/8685496/

[56] Y. Hu, Z. Li, J. Lan, J. Wu, and L. Yao, "EARS: Intelligence-driven experiential network architecture for automatic routing in software-defined networking," *China Communications*, vol. 17, no. 2, pp. 149–162, 2020.

[57] Y. Kim, S. Kim, and H. Lim, "Reinforcement Learning Based Resource Management for Network Slicing," *Applied Sciences*, vol. 9, no. 11, p. 2361, jun 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/11/2361

[58] M. R. Raza, C. Natalino, L. Wosinska, and P. Monti, "Machine Learning Methods for Slice Admission in 5G Networks," *OECC/PSC 2019 - 24th OptoElectronics and Communications Conference/International Conference Photonics in Switching and Computing 2019*, vol. 1, pp. 1–3, 2019.

[59] T. Subramanya, D. Harutyunyan, and R. Riggio, "Machine learning-driven service function chain placement and scaling in MEC-enabled 5G networks," *Computer Networks*, vol. 166, p. 106980, jan 2020.

[60] O. A. Wahab, N. Kara, C. Edstrom, and Y. Lemieux, "MAPLE: A Machine Learning Approach for Efficient Placement and Adjustment of Virtual Network Functions," *Journal of Network and Computer Applications*, vol. 142, pp. 37–50, sep 2019.

[61] G. Zhu, J. Zan, Y. Yang, and X. Qi, "A Supervised Learning Based QoS Assurance Architecture for 5G Networks," *IEEE Access*, vol. 7, no. c, pp. 43 598–43 606, 2019.

[62] S. Mccanne and V. Jacobson, "The BSD Packet Filter: A New Architecture for User-level Packet Capture," 1992.

[63] T. Høiland-Jørgensen, J. D. Brouer, D. Borkmann, J. Fastabend, T. Herbert, D. Ahern, and D. Miller, "The eXpress data path: Fast programmable packet processing in the operating system kernel," *CoNEXT 2018 - Proceedings of the 14th International Conference on Emerging*

Networking EXperiments and Technologies, vol. 18, pp. 54–66, dec 2018. [Online]. Available: https://dl.acm.org/doi/10.1145/3281411.3281443

[64] M. Bertrone, S. Miano, F. Risso, and M. Tumolo, "Accelerating linux security with eBPF iptables," *SIGCOMM 2018 - Proceedings of the 2018 Posters and Demos, Part of SIGCOMM 2018*, pp. 108–110, aug 2018. [Online]. Available: https://dl.acm.org/doi/10.1145/3234200.3234228

[65] S. Rivera, V. K. Gurbani, S. Lagraa, A. K. Iannillo, and R. State, "Leveraging eBPF to preserve user privacy for DNS, DoT, and DoH queries," *ACM International Conference Proceeding Series*, aug 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3407023.3407041

[66] X. Dong and Z. Liu, "Multi-dimensional detection of Linux network congestion based on eBPF," *Proceedings - 2022 14th International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2022*, pp. 925–930, 2022.

[67] H. ElBouanani, C. Barakat, W. Dabbous, and T. Turletti, "Passive delay measurement for fidelity monitoring of distributed network emulation," *Computer Communications*, vol. 195, pp. 40–48, nov 2022.

[68] T. A. Do Amaral, R. V. Rosa, D. F. Moura, and C. E. Rothenberg, "Run-Time Adaptive In-Kernel BPF/XDP Solution for 5G UPF," *Electronics 2022, Vol. 11, Page 1022*, vol. 11, no. 7, p. 1022, mar 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/7/1022/htmhttps://www.mdpi.com/2079-9292/11/7/1022

[69] P. Salva-Garcia, R. Ricart-Sanchez, E. Chirivella-Perez, Q. Wang, and J. M. Alcaraz-Calero, "XDP-Based SmartNIC Hardware Performance Acceleration for Next-Generation Networks," *Journal of Network and Systems Management*, vol. 30, no. 4, pp. 1–26, oct 2022. [Online]. Available: https://link.springer.com/article/10.1007/s10922-022-09687-z

[70] M. S. Brunella, G. Belocchi, G. Bianchi, L. Petrucci, S. Pontarelli, A. Cammarano, M. Bonola, A. Palumbo, G. Siracusano, and R. Bifulco, "hXDP," *Communications of the ACM*, vol. 65, no. 8, pp. 92–100, jul 2022. [Online]. Available: https://dl.acm.org/doi/10.1145/3543668

[71] F. Wang, G. Zhao, Q. Zhang, H. Xu, W. Yue, and L. Xie, "OXDP: Offloading XDP to SmartNIC for Accelerating Packet Processing," *Proceedings of the International Conference on Parallel and Distributed Systems - ICPADS*, vol. 2023-January, pp. 754–761, 2023.

[72] P4.org, "P4-16 Language Specification," 2021. [Online]. Available: https://p4.org/p4-spec/docs/P4-16-v1.2.2.html

[73] P. Bosshart, D. Daly, G. Gibb, M. Izzard, N. McKeown, J. Rexford, C. Schlesinger, D. Talayco, A. Vahdat, G. Varghese, and D. Walker, "P4: programming protocol-independent packet processors," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 3, pp. 87–95, jul 2014. [Online]. Available: https://doi.org/10.1145/2656877.2656890

[74] E. F. Kfoury, J. Crichigno, and E. Bou-Harb, "An Exhaustive Survey on P4 Programmable Data Plane Switches: Taxonomy, Applications, Challenges, and Future Trends," *IEEE Access*, vol. 9, pp. 87 094–87 155, 2021.

[75] S. Aalibagi, M. Dolati, S. Sadrhaghighi, and M. Ghaderi, "Low-Overhead Packet Loss Diagnosis for Virtual Private Clouds using P4-Programmable NICs," *Proceedings of IEEE/IFIP Network Operations and Management Symposium 2023, NOMS 2023*, 2023.

[76] M. Patetta, S. Secci, and S. Taktak, "A Lightweight Southbound Interface for Standalone P4-NetFPGA SmartNICs," *2022 1st International Conference on 6G Networking, 6GNet 2022*, 2022.

[77] J. M. Tilli and R. Kantola, "Data plane protocols and fragmentation for 5G," *2017 IEEE Conference on Standards for Communications and Networking, CSCN 2017*, pp. 207–213, oct 2017.

[78] S. K. Singh, C. E. Rothenberg, G. Patra, and G. Pongracz, "Offloading Virtual Evolved Packet Gateway User Plane Functions to a Programmable ASIC," in *Proceedings of the 1st ACM CoNEXT Workshop on Emerging In-Network Computing Paradigms*, ser. ENCP '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 9–14. [Online]. Available: https://doi.org/10.1145/3359993.3366645

[79] R. MacDavid, C. Cascone, P. Lin, B. Padmanabhan, A. ThakuR, L. Peterson, J. Rexford, and O. Sunay, "A P4-based 5G User Plane Function," in *Proceedings of the ACM SIGCOMM Symposium on SDN Research (SOSR)*, ser. SOSR '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 162–168. [Online]. Available: https://doi.org/10.1145/3482898.3483358

[80] F. Paolucci, F. Cugini, P. Castoldi, and T. Osinski, "Enhancing 5G SDN/NFV Edge with P4 Data Plane Programmability," *IEEE Network*, vol. 35, no. 3, pp. 154–160, may 2021.

[81] F. Paolucci, D. Scano, F. Cugini, A. Sgambelluri, L. Valcarenghi, C. Cavazzoni, G. Ferraris, and P. Castoldi, "User Plane Function Offloading in P4 switches for enhanced 5G Mobile Edge Computing," *2021 17th International Conference on the Design of Reliable Communication Networks, DRCN 2021*, apr 2021.

[82] S. K. Singh, C. E. Rothenberg, J. Langlet, A. Kassler, P. Voros, S. Laki, and G. Pongracz, "Hybrid P4 Programmable Pipelines for 5G gNodeB and User Plane Functions," *IEEE Transactions on Mobile Computing*, vol. 22, no. 12, pp. 6921–6937, dec 2023.

[83] R. Kamran, S. Kiran, P. Jha, A. Karandikar't, and P. Chaporkar, "Green 6g: Energy awareness in design," *2024 16th International Conference on COMmunication Systems and NETworkS, COMSNETS 2024*, pp. 1122–1125, 2024.

[84] C. Hardegen, B. Pfulb, S. Rieger, and A. Gepperth, "Predicting Network Flow Characteristics Using Deep Learning and Real-World Network Traffic," *IEEE Transactions on Network and Service Management*, vol. 17, no. 4, pp. 2662–2676, dec 2020.

[85] J. Zhou, X. Qiu, Z. Li, G. Tyson, Q. Li, J. Duan, and Y. Wang, "Antelope: A Framework for Dynamic Selection of Congestion Control Algorithms," *Proceedings - International Conference on Network Protocols, ICNP*, vol. 2021-November, 2021.

[86] X. Zhang, Z. Liu, and J. Bai, "Linux Network Situation Prediction Model Based on eBPF and LSTM," *International Conference on Intelligent Systems and Knowledge Engineering*, pp. 551–556, 2021.

[87] C. Liu, Z. Cai, B. Wang, Z. Tang, and J. Liu, "A protocol-independent container network observability analysis system based on eBPF," *Proceedings of the International Conference on Parallel and Distributed Systems - ICPADS*, vol. 2020-December, pp. 697–702, dec 2020.

[88] J. Yang, L. Chen, and J. Bai, "Redis automatic performance tuning based on eBPF," *Proceedings - 2022 14th International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2022*, pp. 671–676, 2022.

[89] H. Chang, M. Kodialam, T. V. Lakshman, and S. Mukherjee, "Microservice fingerprinting and classification using machine learning," *Proceedings - International Conference on Network Protocols, ICNP*, vol. 2019-October, oct 2019.

[90] M. Bachl, J. Fabini, and T. Zseby, "A flow-based IDS using Machine Learning in eBPF," feb 2021. [Online]. Available: https://arxiv.org/abs/2102.09980v3

[91] S. Y. Wang and J. C. Chang, "Design and implementation of an intrusion detection system by using Extended BPF in the Linux kernel," *Journal of Network and Computer Applications*, vol. 198, p. 103283, feb 2022.

[92] S. Miano, M. Bertrone, F. Risso, M. Tumolo, and M. V. Bernal, "Creating complex network services with eBPF: Experience and lessons learned," *IEEE International Conference on High Performance Switching and Routing, HPSR*, vol. 2018-June, jun 2018.

[93] M. Jadin, Q. De Coninck, L. Navarre, M. Schapira, and O. Bonaventure, "Leveraging eBPF to Make TCP Path-Aware," *IEEE Transactions on Network and Service Management*, vol. 19, no. 3, pp. 2827–2838, sep 2022.

[94] W. Saad, M. Bennis, and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, may 2020.

[95] Z. Qadir, K. N. Le, N. Saeed, and H. S. Munawar, "Towards 6G Internet of Things: Recent advances, use cases, and open challenges," *ICT Express*, vol. 9, no. 3, pp. 296–312, jun 2023.

[96] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, D. Niyato, O. Dobre, and H. V. Poor, "6G Internet of Things: A Comprehensive Survey," *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 359–383, jan 2022.

[97] 3GPP, "3GPP TR 23.287 v17.4.0: Architecture enhancements for 5g system (5gs) to support vehicle-to-everything (v2x) services (release 17)," 3rd Generation Partnership Project, 3GPP Standard, September 2022.

[98] 3GPP, "3GPP TR 22.501 v17.6.0: System architecture for the 5g system (5gs); stage 2 (release 17)," 3rd Generation Partnership Project, 3GPP Standard, September 2022.

[99] 3GPP, "3GPP TR 22.502 v17.6.0: Procedures for the 5g system (5gs); stage 2 (release 17)," 3rd Generation Partnership Project, 3GPP Standard, September 2022.

[100] S. A. Abdel Hakeem, H. H. Hussein, and H. W. Kim, "Vision and research directions of 6G technologies and applications," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 2419–2442, jun 2022.

[101] S. Dang, O. Amin, B. Shihada, and M.-S. Alouini, "What should 6G be?" *Nature Electronics*, vol. 3, no. 1, pp. 20–29, jan 2020.

[102] H. H. H. Mahmoud, A. A. Amer, and T. Ismail, "6G: A comprehensive survey on technologies, applications, challenges, and research problems," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 4, apr 2021.

[103] J. R. Bhat and S. A. Alqahtani, "6G Ecosystem: Current Status and Future Perspective," *IEEE Access*, vol. 9, pp. 43 134–43 167, 2021.

[104] B. Ji, Y. Han, S. Liu, F. Tao, G. Zhang, Z. Fu, and C. Li, "Several Key Technologies for 6G: Challenges and Opportunities," *IEEE Communications Standards Magazine*, vol. 5, no. 2, pp. 44–51, jun 2021.

[105] A. I. Salameh and M. El Tarhuni, "From 5G to 6G—Challenges, Technologies, and Applications," *Future Internet*, vol. 14, no. 4, p. 117, apr 2022.

[106] E. Yaacoub and M.-S. Alouini, "A Key 6G Challenge and Opportunity—Connecting the Base of the Pyramid: A Survey on Rural Connectivity," *Proceedings of the IEEE*, vol. 108, no. 4, pp. 533–582, apr 2020.

[107] M. Vaezi, A. Azari, S. R. Khosravirad, M. Shirvanimoghaddam, M. M. Azari, D. Chasaki, and P. Popovski, "Cellular, Wide-Area, and Non-Terrestrial IoT: A Survey on 5G Advances and the Road Toward 6G," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 1117–1174, 2022.

[108] A. L. Imoize, O. Adedeji, N. Tandiya, and S. Shetty, "6G Enabled Smart Infrastructure for Sustainable Society: Opportunities, Challenges, and Research Roadmap," *Sensors*, vol. 21, no. 5, p. 1709, mar 2021.

[109] C. Yeh, G. D. Jo, Y.-J. Ko, and H. K. Chung, "Perspectives on 6G wireless communications," *ICT Express*, jan 2022.

[110] A. Mourad, R. Yang, P. H. Lehne, and A. de la Oliva, "Towards 6g: Evolution of key performance indicators and technology trends," in *2020 2nd 6G Wireless Summit (6G SUMMIT)*, 2020, pp. 1–5.

[111] H. Guo, X. Zhou, J. Liu, and Y. Zhang, "Vehicular intelligence in 6g: Networking, communications, and computing," *Vehicular Communications*, vol. 33, p. 100399, 2022.

[112] M. Noor-A-Rahim, Z. Liu, H. Lee, M. O. Khyam, J. He, D. Pesch, K. Moessner, W. Saad, and H. V. Poor, "6g for vehicle-to-everything (v2x) communications: Enabling technologies, challenges, and opportunities," *Proceedings of the IEEE*, vol. 110, no. 6, pp. 712–734, 2022.

[113] 3GPP, "3GPP TR 22.186 v16.2.0: Service requirements for enhanced V2X scenarios (release 16)," 3rd Generation Partnership Project, 3GPP Standard, November 2021.

[114] D. P. Moya Osorio, I. Ahmad, J. D. V. Sánchez, A. Gurtov, J. Scholliers, M. Kutila, and P. Porambage, "Towards 6g-enabled internet of vehicles: Security and privacy," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 82–105, 2022.

5.2. Thesis compendium publications

[115] J. Gallego-Madrid, R. Sanchez-Iborra, P. M. Ruiz, and A. F. Skarmeta, "Machine learning-based zero-touch network and service management: a survey," *Digital Communications and Networks*, vol. 8, no. 2, pp. 105–123, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352864821000614

[116] J. Gallego-Madrid, A. Molina-Zarca, R. Sanchez-Iborra, J. Ortiz, and A. F. Skarmeta, "Fast traffic processing in multi-tenant 5g environments: A comparative performance evaluation of p4 and ebpf technologies," *Engineering Science and Technology, an International Journal*, vol. 52, p. 101678, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2215098624000648

[117] J. Gallego-Madrid, I. Bru-Santa, A. Ruiz-Rodenas, R. Sanchez-Iborra, and A. Skarmeta, "Machine learning-powered traffic processing in commodity hardware with ebpf," *Computer Networks*, vol. 243, p. 110295, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128624001270

[118] J. Gallego-Madrid, R. Sanchez-Iborra, J. Ortiz, and J. Santa, "The role of vehicular applications in the design of future 6g infrastructures," *ICT Express*, vol. 9, no. 4, pp. 556–570, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405959523000383

5.3. Other thesis-related publications

[119] J. Sanchez-Gomez, J. Gallego-Madrid, R. Sanchez-Iborra, J. Santa, and A. F. Skarmeta, "Impact of schc compression and fragmentation in lpwan: A case study with lorawan," *Sensors*, vol. 20, no. 1, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/1/280

[120] J. Gallego-Madrid, A. Molina-Zarca, R. Sanchez-Iborra, J. Bernal-Bernabe, J. Santa, P. M. Ruiz, and A. F. Skarmeta-Gómez, "Enhancing extensive and remote lora deployments through mec-powered drone gateways," *Sensors 20(15):4109*, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/15/4109

[121] J. Gallego-Madrid, R. Sanchez-Iborra, J. Santa, and A. Skarmeta, "Evaluation of a zone encryption scheme for vehicular networks," *Computer Networks*, vol. 182, p. 107523, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S138912862031183X

[122] J. Gallego-Madrid, A. Hermosilla, and A. Skarmeta, "5gasp: Security and trust in netapp deployment and operation," in *2021 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, 2021.

[123] K. Trantzas, C. Tranoris, S. Denazis, R. Direito, D. Gomes, J. Gallego-Madrid, A. Hermosilla, and A. Skarmeta, "An automated ci/cd process for testing and deployment of network applications over 5g infrastructure," in *2021 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*, 2021, pp. 156–161.

[124] J. Gallego-Madrid, R. Sanchez-Iborra, and A. Skarmeta, "Security and trust in the integration of network functions within the 5g architecture: The 5gasp project," in *MobiSec 2022 : The 6th International Conference on Mobile Internet Security*, 2022.

[125] J. Gallego-Madrid, R. Sanchez-Iborra, and A. Skarmeta, "From network functions to netapps: The 5gasp methodology," *Computers, Materials & Continua*, vol. 71, no. 2, pp. 4115–4134, 2022. [Online]. Available: http://www.techscience.com/cmc/v71n2/45796

[126] J. Gallego-Madrid, L. Bernal-Escobedo, R. Asensio, A. Hermosilla, A. M. Zarca, J. Ortiz, R. Sanchez-Iborra, and A. Skarmeta, "Gaia 5g: A multi-access smart-campus architecture," in *Internet of Things*, A. González-Vidal, A. Mohamed Abdelgawad, E. Sabir, S. Ziegler, and L. Ladid, Eds. Cham: Springer International Publishing, 2022, pp. 363–374.

[127] K. Trantzas, C. Tranoris, S. Denazis, R. Direito, D. Gomes, J. Gallego-Madrid, A. Hermosilla, and A. Skarmeta, "Implementing a holistic approach to facilitate the onboarding, deployment and validation of netapps," in *2022 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*, 2022, pp. 261–267.

[128] A. Hermosilla, J. Gallego-Madrid, P. Martinez-Julia, V. Kafle, K. Trantzas, C. Tranoris, R. Direito, D. Gomes, J. Ortiz, S. Denazis, and A. Skarmeta, "Deployment of 5g network applications over multidomain and dynamic platforms," in *2022 IEEE Future Networks World Forum (FNWF)*, 2022, pp. 276–281.

[129] J. Gallego-Madrid, A. Hermosilla, and A. F. Skarmeta, "Dynamic deployment and testing of virtual on-board units in 5g," in *2022 IEEE Future Networks World Forum (FNWF)*, 2022, pp. 305–309.

[130] J. Gallego-Madrid, I. Bru-Santa, R. Sanchez-Iborra, and A. Skarmeta, "Procesamiento de tráfico en el kernel de linux con machine learning," in *XV Jornadas de Ingeniería Telemática (JITEL 2023)*, 2023.

[131] J. Gallego-Madrid, A. Hermosilla, L. Bernal-Escobedo, R. Asensio-Garriga, A. Pogo-Medina, J. S. Diez-Revenga, R. Sanchez-Iborra, and A. Skarmeta, "Virtual on-board unit migration in a multi-access smart-campus 5g architecture," in *2023 IEEE Future Networks World Forum (FNWF)*, 2023, pp. 1–6.

[132] J. Gallego-Madrid, I. Bru-Santa, R. Sanchez-Iborra, and A. Skarmeta, "Integrating machine learning models into the linux kernel: Opportunities and challenges," in *MobiSec 2023 : The 7th International Conference on Mobile Internet Security*, 2023.

[133] J. Gallego-Madrid, R. Sanchez-Iborra, and A. S. Gomez, *eBPF and XDP Technologies as Enablers for Ultra-Fast and Programmable Next-Gen Network Infrastructures.* Singapore: Springer Nature Singapore, 2024, pp. 269–283. [Online]. Available: https://doi.org/10.1007/978-981-97-2644-8_13