# Investigación sobre optimización de multicast routing para redes 6G

Agustín Rebechi

Universidad Nacional de San Martín

Docente: Oscar Filevich

## I. Introducción

Mientras todavía estamos adaptándonos al 5G, los grandes laboratorios y empresas tecnológicas ya están trabajando en algo mucho más ambicioso. Las redes 6G, una tecnología que permite velocidades de hasta 1 Tbps, esto serían velocidades 50 veces superiores al 5G, latencias sub milisegundo y con proyección de capacidad de conectar millones de dispositivos en un solo kilómetro cuadrado.

La aparición de las redes 6G está impulsando un cambio de paradigma en los servicios multimedia, permitiendo aplicaciones que requieren gran ancho de banda como por ejemplo streaming, telepresencia holográfica (que requiere una densidad de tráfico de 1-10Tbps/Km²), el video volumétrico en tiempo real (Con velocidades máximas de mas de 100Gbps), la realidad extendida multisensorial XR, coordinación de vehículos autónomos e industria automatizada. Esto va de la mano de un crecimiento explosivo del tráfico multimedia global, y se prevé que solo el mercado del streaming crezca de $87.3 billion en 2023 a $3.78 trillion de dólares para 2030.

Estas transformaciones de servicios imponen retos sin precedentes a la infraestructura de red. La necesidad de soportar demandas heterogéneas de Quality of Service (QoS) se enfrenta a la restricción operativa crítica de minimizar el flujo de red para reducir el consumo de energía y los costes de infraestructura en topologías 6G que cada vez son más complejas. La capacidad de entregar datos frescos y sincronizados a múltiples puntos de destino simultáneamente no es solo una ventaja técnica: es un requisito fundamental para habilitar tecnologías de próxima generación.

Los tradicionales caminos mínimos, tal como Dijkstra o Bellman-Ford; son eficientes en identificar las rutas de costo mínimo entre un único punto de partida y un único punto de llegada en tiempo polinomial. Sin embargo, en escenarios donde se envuelven múltiples usuarios a la vez requiriendo el mismo video stream, estos algoritmos revelan ineficiencia, generan flujo redundante y suelen fallar al provisionar demandas heterogéneas, llevando al uso subóptimo de recursos.

El problema de construir un árbol de mínima transmisión que conecte una fuente con múltiples destinos puede modelarse como una variante del Árbol de Steiner**.** Este problema es ampliamente

conocido en teoría de grafos por ser NP-hard, lo que implica que, en general, no existen algoritmos eficientes que garanticen el óptimo para instancias grandes. En contextos de red reales, los algoritmos exactos se vuelven impracticables por su explosión combinatoria: evaluar todas las posibles elecciones de nodos Steiner o todas las configuraciones de subárboles rápidamente supera cualquier límite razonable de tiempo. Además, su dependencia del conocimiento de la topología global y del cálculo iterativo los hace inadecuados para entornos periféricos distribuidos y con limitaciones de latencia. Peor aún, estos tradicionales algoritmos de multicast asume homogeneidad de demanda

Es por eso que surgieron dos líneas de soluciones prácticas. Una de ellas consiste en algoritmos de aproximación y heurísticas basadas en Steiner, con limitaciones con respecto a adaptarse a diversidad de usuarios y dinámica topológica y asumen homogeneidad en la demanda. Protocolos como el Distance Vector Multicast Routing Protocol (DVMRP), Multicast OSPF (MOSPF), y Protocol Independent Multicast (PIM) surgieron como métodos fundamentales en multicast routing. La segunda línea es más reciente y se apoya en modelos de aprendizaje, y de redes neuronales, aunque ofrecen una mayor velocidad de inferencia, suelen carecer de escalabilidad y generalización topológicas. Sin embargo, a pesar de su potencia, los modelos enrutamiento basados en redes neuronales tradicionales (NN) tiene limitaciones críticas cuando es implementado en escenarios realistas 6G. Uno de los mayores obstáculos: falta de escalabilidad estructural. MLC y CNN requieren dimensiones de entrada y salida fijas, lo que las limita a redes con un número predeterminado de nodos y enlaces. Cualquier cambio en el número de usuarios o modificación en la topología de la red requiere volver a entrar el modela o rediseñar la arquitectura. Dado que su aprendizaje se basa a menudo en información posicional o indexación absoluta, su capacidad de inferencia se degradan cuando se presenta una red que difiere de los datos de entrenamiento. Existe un consenso cada vez mayor en que los algoritmos de enrutamiento no solo deben basarse en datos, sino que también deben ser sensibles a los grafos, adaptables a la topología y escalables.

Es por eso que en este paper voy a analizar y cruzar tres papers recientes, con menos de 2 años de antigüedad. Desde sus propios enfoques, algunos más tradicionales como algoritmos basados en árboles de steiner sobre demanda y programación dinámica, hasta redes neuronales basada en grafos (GNNs) con aprendizaje por refuerzo.

## II Desarrollo

### II-A. On-Demand Steiner Tree (OST)

En este paper se aborda un reto fundamental sin resolver: el problema de flujo mínimo (MFP) con demandas de salida heterogéneas y multidestino, fundamental para la distribución multimedia eficiente, como la transmisión de video con resolución adaptativa. Para superar las limitaciones, propone un algoritmo de árbol de steiner bajo demanda (OST) mejorado con programación dinámica en dos etapas: el primer enfoque que optimiza conjuntamente la agregación de flujo y la selección de rutas sensibles a la QoS

En escenarios multimedia similares a 6G muestran que el OST reduce el flujo total de la red en más de un 10% en comparación con métodos más avanzados, al tiempo que garantiza QoS

El algoritmo On-demand Steiner Tree (OST) replantea de forma radical la optimización del flujo multicast. Introduce un mecanismo de generación de árboles adaptativo que emplea un enfoque de programación dinámica de dos etapas: En primer lugar, descompone la red en subgrafos homogéneos en cuanto a Quality of Service (QoS) a través de la agrupación de requisitos de salida y, a continuación , conecta de forma óptima estos subgrafos a través de uniones de Steiner de coste mínimo. Las principales contribuciones de este paper se resumen a continuación:

1) Establece el primer marco formal introduciendo el problema de flujo mínimo con restricción de salida (OCMFP) como una nueva formulación matemática que capture el equilibrio entre la prestación de servicios QoS y eficiencia en la red
2) Es propuesto un OST que es dinámicamente programable que no solo garantiza mínimo flujo, sino también alcanza adaptabilidad gracias a su mecanismo de construcción de rutas en dos etapas.
3) Los resultados de los experimentos demuestran que el método OST propuesto reduce el tráfico de red en más de un 10% en comparación con los algoritmos tradicionales de optimización de flujo

Output-Constrained Minimum Flow Problem (OCMFP) generaliza el clásico árbol de Steiner al considerar demandas heterogéneas: cada destino puede requerir diferente cantidad de flujo (por ejemplo, usuarios con video 4K vs 8K). El objetivo es minimizar el flujo total en la red sujeto a restricciones que garantizan la conservación de flujo en nodos intermedios (lo que entra no puede ser menor a lo que sale), satisfacción de demandas individuales en cada destino y por último la estructura de árbol desde la fuente hacia los destinos.

Para resolver el OCMFP eficientemente, el OST emplea programación dinámica en dos etapas. El algoritmo divide el conjunto de destinos en subconjuntos y calcula recursivamente el costo mínimo de transmisión, considerando que cuando se reutilizan rutas, la carga mayor domina el costo total. La complejidad temporal y espacial escalan exponencialmente con el número de destinos $O(V_d)$, lo que limita su aplicabilidad práctica en redes con muchos usuarios.

Aplicación del algoritmo OST:

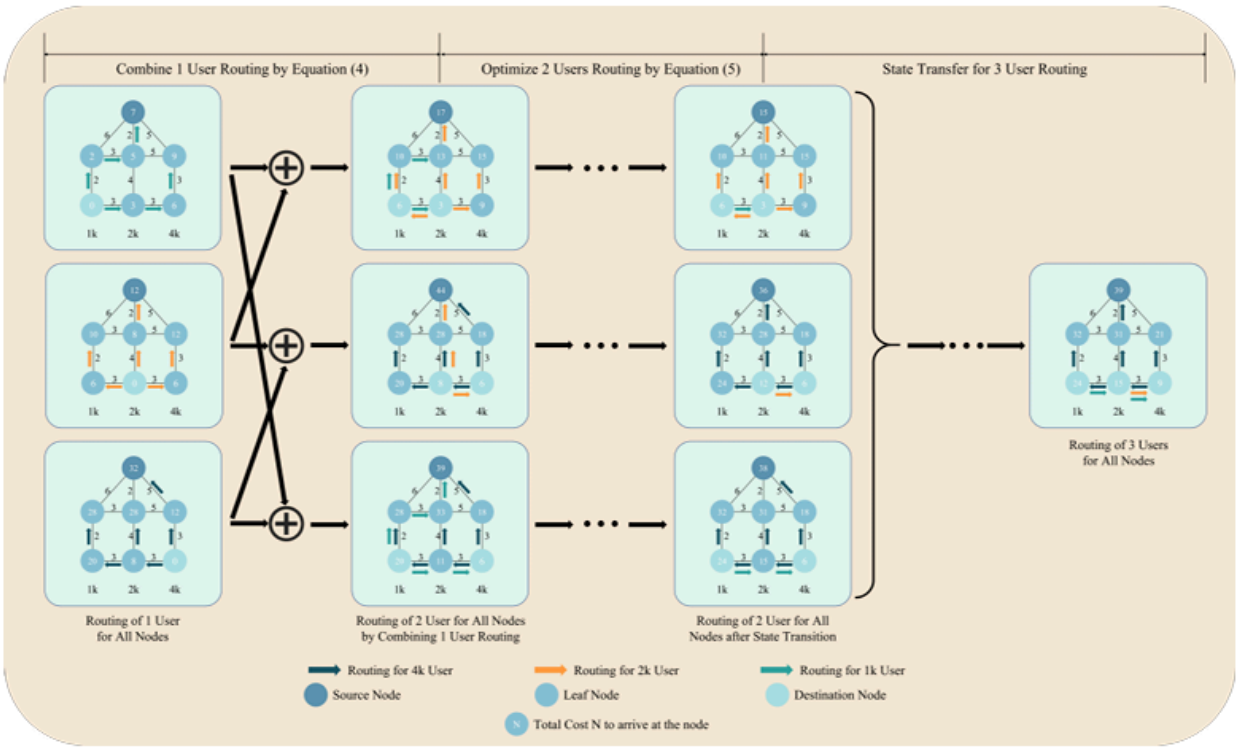



Fig. 1. Application of OST algorithm.

El paper propone dos algoritmos:

Algoritmo 1 (Shortest Video Transmission Path): Encuentra rutas desde un nodo hacia un conjunto de destinos, calculando el costo mínimo para cada subconjunto.

Algoritmo 2 (Optimize Video Transmission Paths): Utiliza programación dinámica con enumeración de subconjuntos de destinos para construir el árbol multicast global óptimo. Este es el algoritmo global que usa DP + enumeración de subconjuntos de destinos. El paper señala que tanto el tiempo como el espacio de este algoritmo escalan exponencialmente en el número de destino $|V_d|$. Lo que lleva a $O(|V| * 2^{V_d})$

## II-B. Graph Neural Network-based (GNN)

Con la evolución de las redes programables, las redes definidas por software (SDN) dan arquitectura centralizada para gestionar multicast más flexiblemente. Los controladores SDN pueden calcular árboles de multicast óptimos o casi óptimos en tiempo real y ajustarlos dinámicamente en función del tráfico y las condiciones de la red.

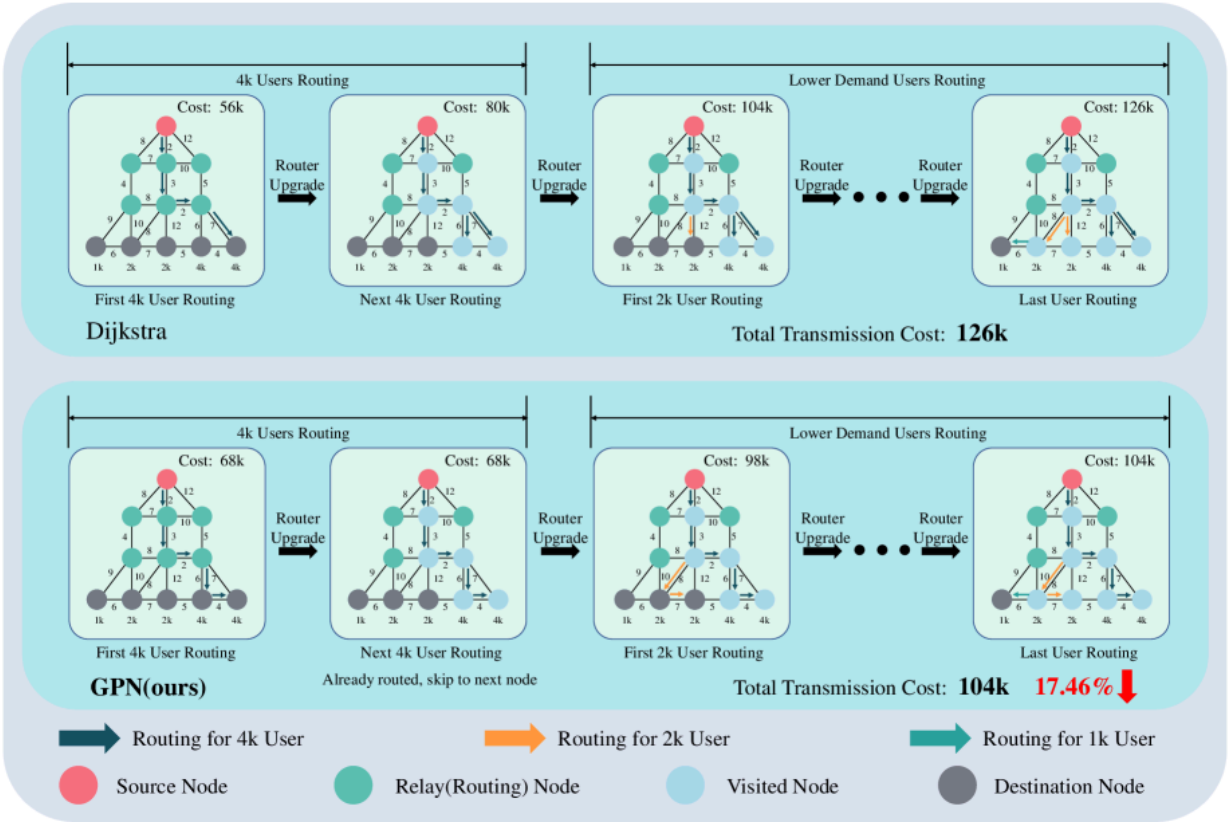
Enfoques basados en redes neuronales, aunque ofrecen una mayor velocidad de inferencia, suelen carecer de escalabilidad y generalización topológica (capacidad de dar respuestas precisas a datos que no vio/no utilizó en entrenamiento). Hay una clara necesidad de nuevos paradigmas de enrutamiento que

sean capaces de: adaptarse a la diversidad de los usuarios, la dinámica topológica y la manejabilidad computacional.

Para abordar estas limitaciones, este paper presenta un marco de enrutamiento multicast basado en GNN que logra minimizar conjuntamente el costo de transmisión total y admite los requisitos de calidad de video específicos del usuario (4k/8k). El problema de enrutamiento se formula como una tarea de Minimum-flow Optimization y se desarrolla un algoritmo de aprendizaje por refuerzo para construir secuencialmente árboles multicast eficientes reutilizando rutas y adaptándose a la dinámica de la red. El algoritmo logra una complejidad de $O(n)$ aprovechando el paradigma de pasos de mensajes.

Una de las características más destacadas de las GNN es el mecanismo de paso de mensajes, que permite a cada nodo intercambiar y agregar información de forma iterativa de sus vecinos. Esta propagación de información local no solo refleja la estructura de muchos algoritmos de enrutamiento distribuidos, sino que también permite a las GNN aprender las representaciones eficientes tanto del contexto de la red global como de las preferencias de ruta local.

Para aprender decisiones de enrutamiento eficientes bajo la formulación secuencial y sensible a la demanda de multicast. Propone el paper una red de políticas grafos (GPN) basada en un codificador de red de atención gráfica (GAT) y un agregador de historial de rutas basado en LSTM. Un Graph Attention Network (GAT) es implementado como encoder para extraer incrustaciones de nodos sensibles al contexto, mientras que un módulo de Long-Short-Term Memory (LSTM) modela las dependencias secuenciales en las decisiones del multicast




En la imagen se ve, que aunque con mayor costo inicial, el GPN tiene mayor entendimiento global del enrutamiento y en resultados finales, obtiene menor cantidad de costo de transmisión total.

En este paper, consideramos un problema generalizado de flujo de coste mínimo en el que un único nodo fuente difunde datos a múltiples nodos de destino a través de una red. La red subyacente se modela como un grafo ponderado no dirigido

El problema que resuelve el paper es una generalización del árbol de Steiner: en vez de solo decidir qué enlaces usar, también decide cuánto flujo enviar por cada uno, considerando que cada destino puede necesitar cantidades diferentes. Esto es un problema de optimización NP-hard, y la GNN aprende a aproximar la solución óptima muchísimo mejor que heurísticas clásicas.

Se evaluó s en redes de 30 a 50 nodos con 12 usuarios de demandas heterogéneas. El método propuesto, GPN, logra costos solo 7% superiores al óptimo teórico, pero con velocidad mil veces mayor. Comparado con Dijkstra, que es el método clásico más rápido, GPN reduce el costo en 24% manteniendo tiempo de respuesta sub-segundo. Esto lo hace viable para redes 6G en tiempo real, donde la topología cambia constantemente

## II-C Graph Attention Network para optimización AoI

Mientras la latencia ha sido tradicionalmente una métrica clave en redes de comunicación, es ampliamente reconocido que minimizar por sí solo el delay no garantiza actualización puntual de

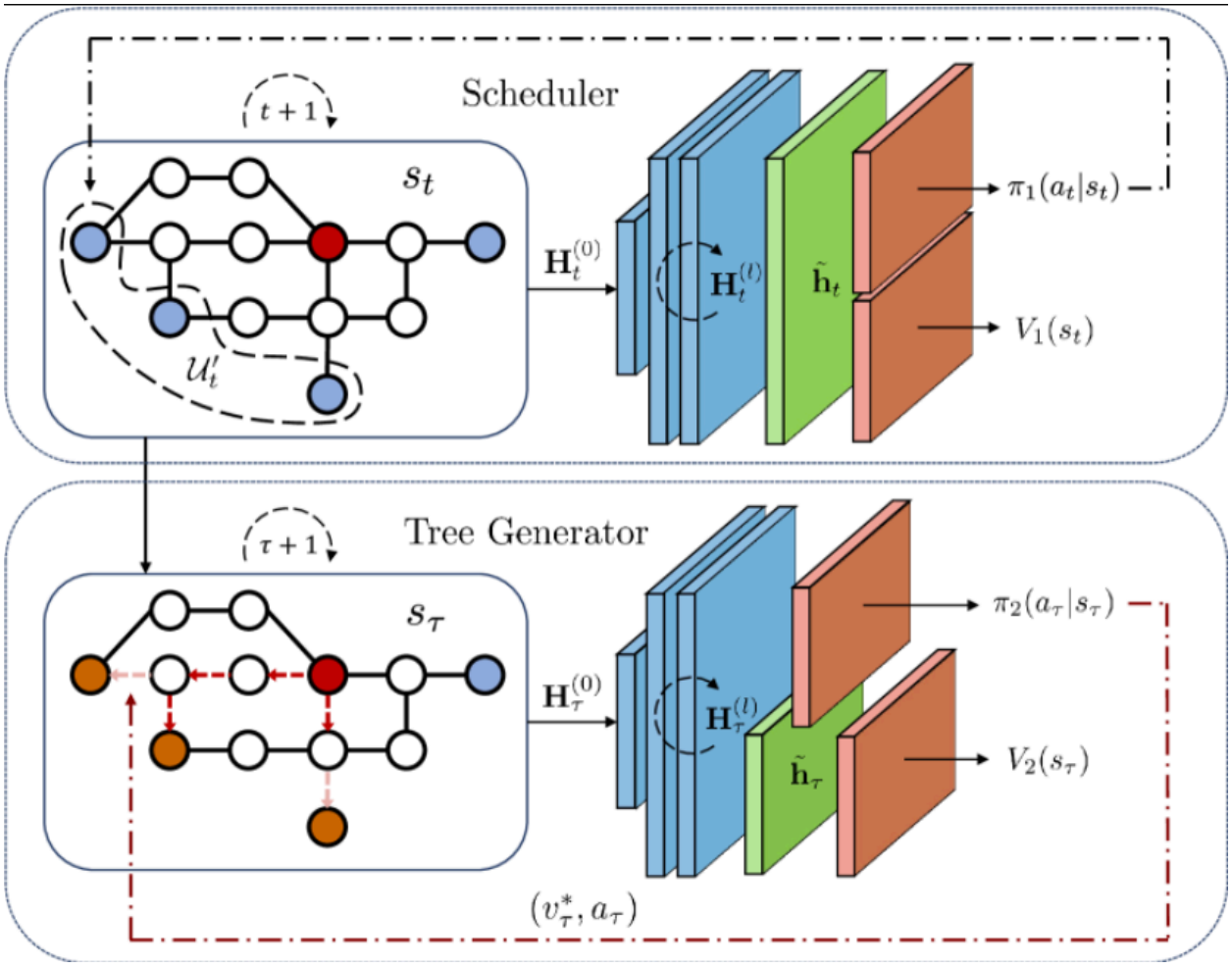
información. Por ejemplo: maximizar la frecuencia de los sensores de actualizaciones puede optimizar la utilización de recursos pero puede llevar a que monitores reciban información desactualizada por message backlogs (atascos) Esto pone de relieve la necesidad de una métrica que capte mejor la oportunidad de la diffusion de la información. El concepto de Age of Information (AoI) ha surgido como una métrica prometedora en campos como el aprendizaje y protocolos de redes. AoI cuantifica la frescura de la información disponible sobre un monitor, lo que hace una métrica de rendimiento adecuada para aplicaciones en tiempo real.

Adicionalmente a diferencia de otras métricas como minimización de costos o delay, la optimización de AoI previene retrasos en actualizaciones y consumo excesivo de recursos, que son críticos para aplicaciones en tiempo real. Aunque algunos estudios han abordado programación multicast para optimización de AoI, a menudo pasan por alto el problema de enrutamiento multicast. Esta limitación surge de la naturaleza de las black box de las estructuras TCP/IP, lo que dificulta compartir información entre capas. Se ha propuesto diseño de capa cruzada (Cross-layers) para abordar estas limitaciones al optimizar conjuntamente las decisiones de enrutamiento y scheduling. En consecuencia un marco multicast de capa cruzada es muy adecuado para optimizar el AoI mediante decisiones conjuntas de enrutamiento y scheduling. Además las redes del mundo real a menudo operan bajo restricciones de energía, donde el consumo total de energía debe gestionarse con cuidado. Esto introduce una complejidad adicional, un trade-off entre el consumo de energía y el AoI.

En primer lugar para resolver conjuntamente el problema de multicast óptimo para AoI de enrutamiento y scheduling (programación), descomponemos el problema original en 2 subproblemas: 1) el subproblema de programación (scheduling) y 2) el subproblema de generación de árboles. Esta descomposición reduce la complejidad del problema original y facilita un diseño por RL Cross-layer. En segundo lugar, la naturaleza NP-hard hace inviable las soluciones óptimas. Proponemos entonces, una heurística de generación de árboles que construye incrementalmente árboles multicast añadiendo nodos y aristas mientras se adhiere a las restricciones de árboles. Tercero, para abordar la alta dimensionalidad del dato en el grafo. Se ha basado en trabajos anteriores que combinan Graph embedding con RL para OC, y han extendido este enfoque a problemas de multicast utilizando Redes de Atención de Grafos (GATs) conocidas por efectividad en el modelado de datos estructurados como grafos con propiedad de contracción

Para lidiar con la complejidad de ambos subproblemas, modelamos un Proceso de Decisión de Markov (MDPs), que con algoritmos RL puede resolver. Como se mencionó, se enfrentan dos desafíos: 1) La dimensionalidad de $MDP_i$ y 2) La complejidad NP_hard de los generadores de árboles del subproblema 2. Primero reformulamos el subproblema de generador de subárboles como un MDP para resolver estos desafíos. Después, se propuso Tree Generator-based Multicast Scheduling (TGMS) Algorithm . Nuestro enfoque propuesto consiste en un generador de árboles y un scheduler, cual utiliza graph embedding metodos y tecnicas DRL. El graph embedding propuesto disminuye la dimensión desde $O(\hat{V}^2)$ a $O(\hat{V})$ con una clave de información extraída. La arquitectura del sistema TGMS se vería de la

siguiente manera:




El scheduler propuesto tiene un rendimiento superior a otros baselines para todos los datasets. reduce el peso promedio de AoI por 21,6% en escenarios de baja energía. También reduce el pico de antigüedad ponderada por 29,9%. El scheduler diseñado es un robusto algoritmo sobre varios topologías basadas en grafos. Los resultados experimentales también demuestran que el esquema es hasta 9.85x veces más eficiente computacionalmente que los algoritmos de enrutamiento multicast tradicionales, al tiempo que logra un rendimiento comparable a métodos SOTA. Adicionalmente, el TGMS supera a otros metodos de referencia, incluidos los diseños no basados en capa cruzada y los métodos sin incrustación de grafo manteniendo un alto nivel de performance y reduciendo AoI promedio ponderado en un 21.1% y la antiguedad maxima ponderada en un 29.7% bajo escenarios de baja energía en multicast

El algoritmo TGMS propuesto tiene algunas limitaciones. En primer lugar, requiere mucha memoria para el entrenamiento de un grafo grande. En segundo lugar, es posible que el árbol multicast no pueda actualizarse en tiempo real debido a la complejidad del grafo. La frecuencia de actualización real depende de los dispositivos SDN utilizados. Una limitación de nuestro algoritmo es que el planificador necesita volver a entrenarse si cambian restricciones de energía. Todo lo anteriormente mencionado motiva potenciales investigaciones futuras

## III. Conclusión

Este trabajo aborda tres enfoques algorítmicos contemporáneos para la optimización de multicast routing para redes 6G. En estos se utilizaron aplicaciones directas de conceptos vistos durante la materia Algoritmos II, tales como: grafos, árboles, recursividad y complejidad computacional. Cabe mencionar que; los árboles y grafos continúan siendo la base teórica fundamental de las redes de comunicación.

Con respecto al algoritmo OST. Al combinar la programación dinámica con árboles de steiner, supera las limitaciones de los enfoques tradicionales que ignoran la multiplexación de flujos o asumen demandas homogéneas. Los resultados experimentales demuestran que reduce en más de un 10 % el flujo de la red en comparación con los métodos existentes, al tiempo que cumple con precisión los diversos requisitos de calidad de servicio. En futuros trabajos se investigará la integración del OST con controladores SDN y su escalabilidad en despliegues de redes ultra densas.

El routing GPN presentó un marco de enrutamiento multicast robusto y escalable que aprovecha las redes neuronales de grafos(GNN) y el aprendizaje por refuerzo para abordar los retos críticos de las demandas heterogéneas de calidad de servicio, el dinamismo de la red y las limitaciones de escalabilidad inherentes a los métodos tradicionales de enrutamiento multicast. Los exhaustivos experimentos realizados subrayan el rendimiento superior a la hora de minimizar los costes de transmisión y mejorar la eficiencia computacional para la implementación en tiempo real de nuestro enfoque, que es muy adecuado para futuras aplicaciones de streaming multimedia dentro del panorama en evolución de las redes 6G.

Con respecto al routing TGMS, formuló una optimización a una métrica tan importante como la latencia: el AoI. Tambien, se ha propuesto una nueva Graph Attention Network (GAT) con la propiedad de contracción para extraer información del grafo. Su diseño de capa cruzada integra enrutamiento y scheduling usando aprendizaje por refuerzo, logrando reducción del 21.6% en AoI promedio bajo restricciones energéticas y eficiencia computacional 9.85x superior a métodos tradicionales.

Comparando los tres papers, se puede observar el siguiente trade-off:  OST garantiza soluciones óptimas pero no escala bien (complejidad exponencial), mientras que GNN y GAT sacrifican optimalidad por velocidad y adaptabilidad. El GPN logró costos solo 7% por encima del óptimo pero mil veces más rápido. Por otro lado, TGMS redujo el AoI en 21.6% con eficiencia 9.85x superior. Esto demuestra que los métodos relacionados al aprendizaje (RL) son más viables para redes 6G en tiempo real.

Este trabajo comenzó con la pregunta sobre cómo distribuir masivas cantidades de datos en las próximas redes 6G, estas deben poder reducir el costo de transmisión, adaptarlas a la diversidad de usuarios, dinámica topología, eficiencia computacional, y bajas latencias; como también AoI. Si bien, las investigaciones actuales todavía tienen algunos límites, el uso de redes neuronales, inteligencia artificial y aprendizaje automático parece ser el camino a seguir de este avance tecnológico que se construye día a día.

La brecha temporal entre estos papers (2024-2025) evidencia la velocidad del avance en este campo, sugiriendo que la convergencia entre teoría de grafos clásica e inteligencia artificial definirá la próxima generación de protocolos de enrutamiento

## Referencias

*Wang, Z., Wang, X., Chang, N., Xu, W., Quan, W., Sun, R., & Zhou, C. (2025). On-demand multimedia delivery in 6G: An optimal-cost Steiner tree algorithm. arXiv. https://arxiv.org/abs/2507.04589*

*Wang, Z., Wang, X., Chang, N., Wenchao, Z., Quan, W., & Shen, X. (2025). Graph neural network-based multicast routing for on-demand streaming services in 6G networks. arXiv. https://arxiv.org/abs/2510.11109*

*Zhang, H., Liao, G., & Cao, S. (2024). Age-minimal multicast by graph attention reinforcement learning. arXiv. https://arxiv.org/abs/2404.18084*

*Latreche, S., Bellahsene, H., & Taleb-Ahmed, A. (2025). Some applications envisaged for the new generation of communications networks 6G. arXiv. https://arxiv.org/abs/2501.14117*