ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ
_____-***_____

BÁO CÁO THU HOẠCH HỌC PHẦN CÁC VẤN ĐỀ HIỆN ĐẠI CỦA TRUYỀN THÔNG VÀ MẠNG MÁY TÍNH

Mô phỏng điều hướng lưu lượng trong O-RAN với xApp dùng Deep Reinforcement Learning

Họ và Tên: Bùi Minh Thắng - 23020646
Nguyễn Vũ Minh - 23020629
Ma Đức Minh - 23020626
Nguyễn Hoàng Tùng Dương - 21020182
Người hướng dẫn: TS. Nguyễn Ngọc Tân

Hà Nội, 2025

Lời cam đoan

Nhóm em xin cam đoan: Báo cáo nghiên cứu khoa học với đề tài "Mô phỏng điều hướng lưu lượng trong O-RAN với xApp dùng Deep Reinforcement Learning" này là của nhóm em. Những gì nhóm em viết ra không có sự sao chép từ các tài liệu, không sử dụng kết quả của người khác mà không trích dẫn cụ thể. Đây là công trình nghiên cứu tập thể nhóm em tự phát triển, không sao chép mã nguồn của người khác. Nếu vi phạm những điều trên, nhóm em xin chấp nhận tất cả những truy cứu về trách nhiệm theo quy định.

Sinh viên

Bùi Minh Thắng
Nguyễn Vũ Minh
Ma Đức Minh
Nguyễn Hoàng Tùng Dương

Lời cảm ơn

Lời đầu tiên, em xin được gửi lời cảm ơn chân thành tới Khoa Công nghệ Thông tin – Trường Đại học Công nghệ – Đại học Quốc gia Hà Nội đã tạo điều kiện thuận lợi để em được học tập, nghiên cứu và thực hiện đề tài này.

Em xin bày tỏ lòng biết ơn sâu sắc tới thầy Nguyễn Ngọc Tân và thầy Nguyễn Thái Dương đã tận tình hướng dẫn, hỗ trợ em trong suốt quá trình nghiên cứu và triển khai đề tài.

Bên cạnh đó, em xin được bày tỏ lòng biết ơn tới các thầy cô trong khoa đã tận tâm giảng dạy và trang bị cho em những kiến thức quý báu trong suốt quá trình học tập tại trường.

Cuối cùng, em xin chúc các thầy cô, các bạn luôn mạnh khỏe, hạnh phúc và gặt hái nhiều thành công trong cuộc sống.

Tóm tắt

Báo cáo này trình bày quá trình phát triển và mô phỏng một xApp trong kiến trúc Open RAN, tập trung vào việc điều hướng lưu lượng mạng sử dụng Deep Reinforcement Learning (DRL). Chúng tôi đã xây dựng một xApp để tối ưu hóa việc phân phối lưu lượng mạng trong môi trường O-RAN, sử dụng mô hình DRL để cải thiện hiệu suất mạng. Bằng cách sử dụng mô phỏng ns-3, chúng tôi đã kiểm tra hiệu quả của xApp trong các tình huống thực tế, cho thấy khả năng cải thiện đáng kể trong việc quản lý lưu lượng và tối ưu hóa tài nguyên mạng. Báo cáo cũng cung cấp hướng dẫn chi tiết về cách phát triển xApp và tích hợp nó vào hệ thống O-RAN.

Từ khoá: Open RAN architecture, O-RAN RIC, xApp development, O-RAN use cases, ns-3 O-RAN simulation, DRL for traffic steering, RAN Intelligent Controller.

Mục lục

Chương 1

Đặt vấn đề

Trong những năm gần đây, sự phát triển bùng nổ của các dịch vụ viễn thông, đặc biệt là các ứng dụng yêu cầu tốc độ cao, độ trễ thấp và khả năng quản lý linh hoạt, đã đặt ra những thách thức mới đối với hạ tầng mạng truyền thống. Kiến trúc mạng truy cập vô tuyến truyền thống (RAN) với các thiết bị và giao diện đóng độc quyền gây ra nhiều hạn chế về tính linh hoạt, khả năng tương tác và đổi mới công nghệ. Chính vì vậy, xu hướng chuyển dịch sang kiến trúc mạng truy cập vô tuyến mở (Open RAN) ngày càng thu hút sự quan tâm lớn từ cộng đồng nghiên cứu và các nhà cung cấp dịch vụ viễn thông hàng đầu thế giới.

Open RAN được kỳ vọng sẽ tái định hình ngành công nghiệp viễn thông nhờ vào việc mở các giao diện, phân tách chức năng mạng (như Central Unit - CU, Distributed Unit - DU, và Radio Unit - RU), đồng thời đưa trí tuệ nhân tạo (AI) vào sâu hơn trong việc quản lý và vận hành mạng lưới. Một thành phần cốt lõi giúp đạt được những lợi ích này chính là bộ điều khiển thông minh mạng vô tuyến (RAN Intelligent Controller - RIC) với hai phiên bản chính là Near-Real-Time RIC và Non-Real-Time RIC. Hai loại RIC này giúp quản lý và tối ưu các ứng dụng mạng (xApps và rApps) trong thời gian thực hoặc gần thực.

Mục tiêu của báo cáo này là làm rõ những ưu điểm của Open RAN so với mạng RAN truyền thống, đồng thời phân tích cụ thể vai trò của các thành phần quan trọng như RIC và các giao diện mở (E2, A1, O1). Qua đó, nhóm nghiên cứu mong muốn đề xuất một mô hình tối ưu hóa tài nguyên mạng hiệu quả bằng công nghệ học tăng cường sâu (DRL) nhằm đáp ứng các yêu cầu về hiệu năng, linh hoạt và khả năng mở rộng của mạng viễn thông thế hệ mới.

Chương 2

Kiến thức cơ sở

2.1 So sánh kiến trúc RAN truyền thống và O-RAN

Các mạng di động truyền thống trong lịch sử chủ yếu áp dụng kiến trúc RAN phân tán (Distributed RAN - D-RAN), trong đó mỗi trạm gốc tích hợp cả chức năng xử lý băng tần gốc (baseband) và chức năng vô tuyến. Mô hình này phù hợp với các thế hệ di động đầu tiên (2G/3G/4G), nhưng tồn tại nhiều hạn chế về khả năng mở rộng, tính tương tác giữa các nhà cung cấp và khả năng quản lý tập trung. Với sự ra đời của 5G và tương lai là 6G, ngành công nghiệp viễn thông đang chuyển dịch sang các giải pháp linh hoạt và định nghĩa bằng phần mềm. Open RAN (O-RAN) đưa ra một kiến trúc mở, mô-đun và tách rời bằng cách phân tách các chức năng RAN thành ba thành phần riêng biệt: CU, DU và RU, cùng với các giao diện tiêu chuẩn hóa [7]. Bài viết này phân tích các điểm khác biệt về kiến trúc và những hàm ý giữa D-RAN và O-RAN.

Trong D-RAN, mỗi trạm thu phát (cell site) chứa đồng thời một Bộ xử lý băng tần gốc (BBU) và một Bộ vô tuyến từ xa (RRU). BBU thực hiện toàn bộ xử lý tầng baseband, bao gồm PHY, MAC, RLC và các tầng giao thức cao hơn, trong khi RRU đảm nhận các chức năng đầu vào/ra RF tương tự. Sự tích hợp này giúp đơn giản hóa việc đồng bộ hóa và giảm độ trễ đường truyền fronthaul, nhưng lại dẫn đến hệ thống cứng nhắc và phụ thuộc vào nhà cung cấp. Do phần cứng và phần mềm gắn kết chặt chẽ, việc mở rộng hoặc nâng cấp hệ thống thường đòi hỏi phải thay thế toàn bộ các thành phần độc quyền. Ngoài ra, khả năng tối ưu hóa mạng và chia sẻ tài nguyên giữa các trạm bị hạn chế do xử lý vẫn diễn ra độc lập tại từng vị trí [1].

O-RAN khác biệt với cách tiếp cận nguyên khối của D-RAN bằng việc phân tách chức năng rõ ràng giữa RU, DU và CU. RU đảm nhận xử lý PHY thấp và RF, DU chịu trách nhiệm xử lý PHY cao, MAC và RLC, trong khi CU điều khiển các tầng SDAP, PDCP và RRC. Các thành phần này giao tiếp thông qua các giao diện mở: giao diện F1 giữa CU và DU, và giao diện fronthaul mở (thường là chia tách 7.2x) giữa DU và RU [4]. Tính mô-đun này cho phép các thiết bị từ nhiều nhà cung cấp khác nhau tương tác, đồng thời hỗ trợ ảo hóa các chức năng CU/DU trên phần cứng thương mại thông dụng [3].

Một điểm khác biệt lớn giữa D-RAN và O-RAN nằm ở tính linh hoạt khi triển khai. Trong khi D-RAN đồng vị tất cả xử lý tại trạm thu phát, thì O-RAN hỗ trợ việc tập trung hóa CU tại các trung tâm dữ liệu khu vực và phân phối DU gần biên mạng. Sự phân tách này cho phép phân bổ tài nguyên động, giảm chi phí đầu tư (CAPEX) và nâng cao khả năng mở rộng. Hơn nữa, O-RAN còn hỗ trợ các chức năng tiên tiến như bộ điều khiển RAN thông minh (RIC), cho phép tối ưu hóa dựa trên AI/ML gần như theo thời gian thực [7].

Tuy có nhiều ưu điểm, O-RAN cũng đặt ra thách thức trong việc tích hợp do môi trường đa nhà cung cấp và yêu cầu nghiêm ngặt về đường truyền fronthaul. Các liên kết có tốc độ cao và độ trễ thấp là điều kiện bắt buộc để đảm bảo hiệu năng trong kiến trúc chia tách. Ngược lại, D-RAN đơn giản hơn trong triển khai và đã được kiểm chứng về độ tin cậy, nhưng thiếu sự linh hoạt và tính mở cần thiết cho các trường hợp sử dụng mới như chia sẻ mạng (network slicing)

và mạng 5G riêng (private 5G) [3].

Kiến trúc D-RAN truyền thống đã đóng vai trò là nền tảng của mạng di động trong nhiều thập kỷ, mang lại sự đơn giản và độ ổn định cho các thế hệ mạng di động trước. Tuy nhiên, nhu cầu ngày càng tăng về tính linh hoạt, hiệu quả chi phí và phân biệt dịch vụ đang thúc đẩy quá trình chuyển đổi sang hệ thống mở và tách rời. O-RAN, với việc phân tách chức năng và các giao diện tiêu chuẩn, đại diện cho một sự chuyển đổi mô hình hướng tới giải pháp RAN gốc đám mây và trung lập với nhà cung cấp. Mặc dù vẫn còn tồn tại các thách thức về tích hợp và tối ưu hóa hiệu năng, nhưng những lợi ích tiềm năng của O-RAN đang định vị nó như một yếu tố then chốt trong việc phát triển mạng di động tương lai [4].

2.2 Chức năng của Near-Real-Time RIC và Non-Real-Time RIC

Trong kiến trúc mạng truy cập vô tuyến mở (Open RAN), Bộ điều khiển RAN thông minh (RAN Intelligent Controller – RIC) được phân tách thành hai phần: RIC thời gian gần thực (Near-Real-Time RIC) và RIC thời gian không thực (Non-Real-Time RIC). Hai thành phần này phối hợp điều khiển và tối ưu mạng ở những quy mô thời gian khác nhau nhằm nâng cao hiệu năng của RAN. Cụ thể, Near-RT RIC đảm nhiệm việc điều khiển RAN với độ trễ thấp (từ khoảng 10 mili-giây đến <1 giây) [5], còn Non-RT RIC phụ trách các tác vụ ở quy mô thời gian dài hơn (>1 giây, thường tính bằng giây, phút hoặc lâu hơn) [5]. Sự phân chia này cho phép tối ưu mạng ở cả thời gian thực ngắn hạn lẫn hoạch định dài hạn, tạo nên hệ thống điều khiển nhiều tầng cho RAN.

Near-Real-Time RIC (Near-RT RIC): Đây là thành phần RIC hoạt động gần thời gian thực, thường được triển khai trên hạ tầng điện toán biên hoặc cụm mạng khu vực gần với các nút RAN. Near-RT RIC có chức năng thu thập thông tin trạng thái mạng và thực thi các hành động điều khiển nhanh lên mạng vô tuyến với độ trễ yêu cầu dưới 1 giây [5]. Theo đặc tả O-RAN, Near-RT RIC là một chức năng logic cho phép điều khiển và tối ưu tài nguyên RAN ở mức độ nhanh, thông qua việc thu thập dữ liệu chi tiết và tác động hành động lên các nút RAN qua giao diện E2 docs.o-ran-sc.org. Near-RT RIC thường xử lý các tác vụ như điều khiển truy cập vô tuyến và tài nguyên vô tuyến theo thời gian thực gần, ví dụ: điều phối lịch truyền dẫn, cân bằng tải giữa các cell, điều chỉnh tham số handover, điều khiển can nhiễu,… nhằm tối ưu hiệu suất thông lượng và chất lượng dịch vụ tức thời cho người dùng. Thành phần này tương tác trực tiếp với các nút mạng RAN (như O-DU, O-CU) qua giao diện E2 để nhận số liệu tình trạng (telemetry) và gửi chỉ thị điều khiển một cách liên tục. Đặc điểm quan trọng của Near-RT RIC là khả năng mở rộng chức năng qua các xApp – những ứng dụng plug-and-play chạy trên nền tảng Near-RT RIC để thực hiện các thuật toán điều khiển radio chuyên biệt [5]. Near-RT RIC cũng có cơ chế phối hợp và tránh xung đột giữa nhiều xApp khác nhau cùng tác động lên RAN (ví dụ: cơ chế quản lý message bus, lớp dữ liệu chia sẻ, và logic phân giải xung đột) để đảm bảo các quyết định điều khiển không mâu thuẫn [5].

Non-Real-Time RIC (Non-RT RIC): Đây là thành phần RIC hoạt động ngoài thời gian thực chặt chẽ, nằm trong khối dịch vụ quản lý và điều hành (Service Management and Orchestration – SMO) ở trung tâm mạng hoặc đám mây. Non-RT RIC chịu trách nhiệm thực hiện các tác vụ quản lý, tối ưu RAN ở quy mô dài hạn hơn (trên 1 giây) [5], bao gồm quản lý chính sách dịch vụ, phân tích hiệu năng, tối ưu cấu hình và hoạch định tài nguyên chiến lược cho mạng. Theo O-RAN Alliance, Non-RT RIC là một chức năng logic trong SMO hỗ trợ điều khiển/tối ưu RAN phi-thời-gian-thực, cung cấp khung AI/ML để huấn luyện và cập nhật mô hình, và truyền tải các hướng dẫn chính sách tới RIC gần thực [2] [8]. Non-RT RIC được cấu thành bởi framework Non-RT RIC (nền tảng) và các ứng dụng rApp (các ứng dụng chạy trên Non-RT RIC). Nền tảng Non-RT RIC thực hiện việc kết thúc (terminate) giao diện A1 với Near-RT RIC, đồng thời

phơi bày dịch vụ quản lý dữ liệu và ML cho các rApp thông qua giao diện nội bộ R1 [8]. Các rApp (RAN applications) là những ứng dụng mô-đun chạy trên Non-RT RIC, sử dụng các dịch vụ mà nền tảng cung cấp để tạo ra các giá trị gia tăng cho vận hành RAN [8]. Nhiệm vụ của rApp rất đa dạng, bao gồm: đề xuất và điều chỉnh chính sách điều khiển RAN, phân tích dữ liệu hiệu năng dài hạn, tối ưu cấu hình tham số, cũng như cung cấp thông tin bổ sung (enrichment information) cho các ứng dụng khác [8]. Non-RT RIC gửi hướng dẫn chính sách và mục tiêu đến Near-RT RIC thông qua giao diện A1 (ví dụ: chính sách về phân bổ tài nguyên, mục tiêu QoS cần đạt, tham số ngưỡng sự kiện, v.v.), nhờ đó ảnh hưởng gián tiếp đến hành vi của các xApp trên Near-RT RIC [8]. Ngược lại, Non-RT RIC cũng thu thập phản hồi từ mạng (thông qua dữ liệu O1 hoặc qua báo cáo từ Near-RT RIC) để đánh giá và điều chỉnh các chiến lược tối ưu. Có thể xem Non-RT RIC như "bộ não" ở tầng trên, vạch ra chiến lược dài hạn cho mạng, trong khi Near-RT RIC là "cánh tay tác động" ở tầng dưới thực thi các điều chỉnh nhanh theo chiến lược đó.

2.3 Giao diện mở (E2, A1, O1) và vai trò của chúng

Giao diện E2 là giao diện mở kết nối trực tiếp bộ điều khiển RAN thông minh thời gian thực gần (Near-RT RIC) với các node mạng RAN (gọi chung là E2-nodes), bao gồm các thành phần như O-CU (khối xử lý tập trung – control/user plane), O-DU (khối xử lý phân tán) và thậm chí cả eNB/gNB nếu tuân thủ O-RAN. Mục đích thiết kế của E2 là cho phép RIC thời gian thực gần thu thập số liệu và trạng thái từ mạng RAN theo thời gian thực, đồng thời gửi các lệnh điều khiển tương ứng xuống các nút mạng này một cách nhanh chóng. Giao diện E2 tạo kênh trao đổi hai chiều: truyền lên các thông tin đo lường, chỉ số hiệu năng, sự kiện (telemetry) từ RAN lên Near-RT RIC, và truyền xuống các lệnh cấu hình, điều khiển từ RIC tới RAN [6]. Với E2, Near-RT RIC (thông qua các xApp chạy trên đó) có thể thực hiện các vòng lặp điều khiển kín gần thời gian thực cho RAN. Cụ thể, E2 hỗ trợ các dịch vụ như giám sát (monitor), tạm dừng hoặc ghi đè (suspend/override) và điều khiển các chức năng trên nút mạng, giúp RIC có thể can thiệp vào hoạt động nội tại của trạm gốc khi cần thiết [6]. Các xApp – là những ứng dụng tác nghiệp thời gian thực gần trên RIC – sử dụng E2 để đăng ký nhận các chỉ số/KPI từ nút mạng và đưa ra quyết định điều khiển tương ứng (ví dụ: điều chỉnh lịch truyền, phân bổ tài nguyên vô tuyến, cân bằng tải, giảm nhiễu) trong khoảng thời gian tính bằng mili-giây đến dưới 1 giây. Nhờ E2, O-RAN cho phép tách rời và mở rộng khả năng quản lý tài nguyên vô tuyến của trạm gốc, tạo điều kiện cho nhiều nhà cung cấp khác nhau tích hợp giải pháp tối ưu riêng vào RAN một cách tiêu chuẩn hóa [6]

Giao diện A1 được định nghĩa để kết nối RIC phi thời gian thực (Non-RT RIC, tích hợp trong khung SMO) với Near-RT RIC. Đây là kênh trao đổi thông tin ở mức chiến lược và chính sách giữa lớp quản lý/tối ưu dài hạn và lớp điều khiển ngắn hạn của RAN. Mục đích thiết kế của A1 là cho phép Non-RT RIC cung cấp định hướng vận hành cho Near-RT RIC dưới dạng các chính sách tối ưu mạng (policy) hoặc thông tin hỗ trợ thông minh. Chẳng hạn, thông qua A1, rApp (ứng dụng chạy trên Non-RT RIC) có thể gửi xuống Near-RT RIC những chính sách điều khiển (ví dụ: ưu tiên phục vụ cho một lát cắt mạng hoặc hạn mức tài nguyên cho một nhóm UE cụ thể) nhằm định hướng hành vi của các xApp trên RIC thời gian thực gần [6]. Chức năng chính của A1 bao gồm truyền tải policy (chính sách điều khiển) và thông tin làm giàu (enrichment information) từ khối Non-RT RIC xuống Near-RT RIC. Các policy A1 được định nghĩa ở mức ý định/cấp cao, ví dụ như mục tiêu QoS hoặc KPI mà hệ thống cần đạt được cho một tập người dùng hoặc một lát cắt mạng [5]. Giao diện A1 cũng được thiết kế để hỗ trợ việc quản lý vòng đời của các mô hình học máy (ML) được sử dụng trong RAN, chẳng hạn như phân phối hoặc cập nhật mô hình ML cho xApp, mặc dù chức năng này vẫn đang được nghiên cứu bổ sung. Ngoài ra, A1 cho phép Non-RT RIC gửi thông tin làm giàu – ví dụ dữ liệu dự đoán từ phân tích dài hạn hoặc thông tin từ nguồn ngoại vi – để hỗ trợ xApp ra quyết định tốt hơn. Ngược lại, Near-RT

RIC có thể phản hồi tối thiểu qua A1 về trạng thái thực thi chính sách (ví dụ chính sách đã được áp dụng hay chưa) để Non-RT RIC theo dõi [5]. Với vai trò cầu nối ở mức phi thời gian thực, A1 đảm bảo rằng những tối ưu dài hạn (từ rApp) được hiện thực hóa kịp thời trong lớp điều khiển RAN ngắn hạn, tạo nên sự kết hợp nhịp nhàng giữa chiến lược tổng thể và tác vụ điều khiển cụ thể trong hệ thống RAN thông minh.

Giao diện O1 là giao diện mở hỗ trợ chức năng quản lý, vận hành và bảo trì (OAM) cho tất cả các thành phần O-RAN, nằm trong khung Quản lý dịch vụ và điều phối (SMO) của hệ thống. Cụ thể, O1 kết nối hệ thống SMO với các phần tử mạng O-RAN (O-CU, O-DU, O-RU, Near-RT RIC, v.v.), cho phép thực hiện các tác vụ FCAPS truyền thống – bao gồm Quản lý lỗi, Cấu hình, Kế toán, Hiệu năng và Bảo mật – trên môi trường RAN mở đa nhà cung cấp [6]. Mục đích thiết kế của giao diện O1 là cung cấp một kênh quản lý tập trung và tiêu chuẩn để cấu hình tham số mạng, phân bổ tài nguyên, giám sát hiệu năng và thu thập dữ liệu thống kê từ các nút mạng khác nhau, bất kể nhà sản xuất. Chức năng chính của O1 bao gồm: quản lý cấu hình (ví dụ: thiết lập cấu hình ban đầu cho O-CU/O-DU/O-RU, điều chỉnh tham số RAN theo lịch hoặc theo yêu cầu tối ưu), giám sát hiệu năng (thu thập KPI, số liệu thống kê, bản ghi sự kiện từ các node mạng), quản lý lỗi (nhận cảnh báo, nhật ký lỗi từ phần tử mạng) và quản lý phần mềm (nâng cấp phần mềm, thay đổi phiên bản cấu kiện mạng) [6]. Trong kiến trúc RAN thông minh, O1 cung cấp dữ liệu nền tảng cho khối Non-RT RIC/rApps phân tích. Chẳng hạn, rApp có thể sử dụng số liệu thu thập qua O1 (như mức tải, công suất, chất lượng kênh từ các cell) để phát hiện xu hướng hoặc bất thường, từ đó đề xuất chính sách tối ưu tương ứng. Đồng thời, O1 cũng thực thi các quyết định quản trị như điều chỉnh cấu hình mạng hoặc bật/tắt tài nguyên vô tuyến khi được yêu cầu (thường là kết quả của các thuật toán tối ưu dài hạn). Nhờ giao diện O1, việc vận hành RAN trở nên linh hoạt và tự động hơn, tạo nền tảng cho quản trị mạng tự động trong môi trường O-RAN.

So sánh và tương tác: Ba giao diện E2, A1, O1 đảm nhiệm các vai trò bổ trợ nhau ở các tầng thời gian khác nhau trong hệ thống RAN thông minh [5]. Thứ nhất, E2 là giao diện thời gian thực gần (độ trễ khoảng 10 ms đến dưới 1 giây) phục vụ cho vòng điều khiển nhanh giữa Near-RT RIC và các nút mạng RAN. Giao diện này cho phép các xApp thu thập tức thời trạng thái mạng và tác động trực tiếp lên tài nguyên vô tuyến (ví dụ điều khiển lịch truyền, phân bổ kênh) nhằm tối ưu cục bộ theo thời gian thực. Thứ hai, A1 hoạt động ở mức phi thời gian thực (độ trễ từ vài giây trở lên), là kênh truyền tải chính sách và thông tin thông minh từ tầng quản lý xuống tầng điều khiển [6]. A1 đảm bảo Near-RT RIC vận hành theo đúng mục tiêu tối ưu dài hạn do Non-RT RIC đề ra (ví dụ đảm bảo thông lượng cell edge, giảm tiêu thụ năng lượng...), hỗ trợ bởi các mô hình ML và phân tích dữ liệu ở tầng trên. Thứ ba, O1 là giao diện quản lý với thời gian phản hồi không đòi hỏi tức thời, chủ yếu phục vụ công tác OAM và cung cấp dữ liệu cho các thuật toán thông minh [5]. O1 cho phép SMO và rApp nhìn bao quát toàn mạng, thu thập số liệu đa miền (radio, truyền dẫn, core) để từ đó tối ưu cấu hình mạng một cách tự động. Trong một chu trình vận hành RAN thông minh, cả ba giao diện kết hợp nhịp nhàng theo mô hình điều khiển phân tầng. Non-RT RIC sử dụng O1 để thu thập số liệu vận hành từ các node RAN (đồng thời nhận thông tin từ core mạng nếu cần), tiến hành phân tích và huấn luyện mô hình ML. Từ kết quả phân tích, rApp sinh ra các chính sách hoặc khuyến nghị tối ưu và gửi xuống Near-RT RIC qua giao diện A1[5]. Near-RT RIC nhận được chỉ dẫn từ A1 sẽ điều chỉnh hành vi của các xApp, cho phép các xApp tác động lên mạng RAN theo thời gian thực gần thông qua giao diện E2 (ví dụ thay đổi tham số lịch trình, hạn chế kết nối vào một cell quá tải, v.v.). Như vậy, E2, A1, O1 tạo thành bộ khung giao diện mở giúp hiện thực hóa RAN mở và thông minh: O1 cung cấp khả năng quan sát và cấu hình toàn diện, A1 truyền tải tri thức và mục tiêu tối ưu, còn E2 thực thi điều khiển linh hoạt tại nút mạng, tất cả đều trên nền tảng tiêu chuẩn chung của O-RAN [6]. Nguồn tài liệu tham khảo: Các thông tin trên được tổng hợp từ đặc tả kỹ thuật của O-RAN Alliance và các nghiên cứu IEEE gần đây. Đặc biệt, định nghĩa và chức năng của các giao diện E2, A1, O1 được mô tả trong tài liệu của O-RAN Alliance [6], cũng như các phân tích học thuật

về kiến trúc O-RAN (như bài báo của Michele Polese và cộng sự) để làm rõ vai trò của chúng trong hệ thống RAN thông minh. Các ví dụ triển khai và use-case minh họa cũng cho thấy sự phối hợp giữa ba giao diện này trong việc hỗ trợ ứng dụng rApp/xApp tối ưu mạng một cách linh hoạt và hiệu quả.

Chương 3

Cài đặt phương pháp và thực nghiệm

3.1 Tổng quan về mô phỏng

Mục tiêu của mô phỏng là đánh giá chiến lược điều hướng lưu lượng **Traffic Steering** trong O-RAN. Mô phỏng bao gồm một gNB trung tâm và bốn gNB góc khi mạng phục vụ hai loại UE: VOICE (lưu lượng nhỏ, ưu tiên QoS thấp-latency) và MBB (mobile broadband — lưu lượng lớn). Mục tiêu chính:

- So sánh hai chính sách phân phối kết nối: **DEFAULT** (kết nối tới gNB gần nhất còn khả dụng) và **OFFLOAD** (VOICE ưu lập góc, MBB ưu trung tâm).
- Đo lường tác động các chính sách lên throughput, latency (thời gian hoàn thành truyền), tình trạng chờ (queue), và phân bổ tải trên gNB.
- Khảo sát trade-off giữa tối ưu hóa cho VOICE vs. MBB, và đánh giá robust của chính sách trước các thay đổi về tham số (dung lượng gNB, tốc độ truyền).

3.2 Mô tả môi trường mô phỏng

3.2.1 Cấu trúc topologi và tham số chính

Mô phỏng diễn ra trên vùng phẳng kích thước $AREA - SIZE = 600 \times 600$. Hệ thống tổ chức gồm:

- **gNBs:** 5 trạm gNB:
  - 1 gNB trung tâm (center) đặt tại tâm với dung lượng = 70 MB.
  - 4 gNB góc (corner) đặt gần các góc vùng với dung lượng = 30 MB mỗi trạm.
- **UEs:** 20 thiết bị:
  - 10 VOICE UEs, mỗi UE có kích thước gói nhỏ (khoảng 3–10 MB).
  - 10 MBB UEs, gói lớn hơn (khoảng 11–25 MB).
- **Tốc độ truyền:** $MB - PER - STEP = 2$ MB mỗi bước mô phỏng (mỗi step tương ứng 1 giây trong mô phỏng).
- **Vùng ưu tiên:** cung cấp bán kính ưu tiên cho policy OFFLOAD:
  - $VOICE - RADIUS = 75$ (UE voice trong bán kính này quanh gNB góc được ưu tiên kết nối).

– $MBB - RADIUS = 150$ (UE MBB trong bán kính này quanh gNB trung tâm được ưu tiên).

3.2.2 Luồng hoạt động cơ bản mô phỏng

Mô phỏng hoạt động theo các bước (step-by-step):

1. Cập nhật tải hiện tại trên từng gNB theo các kết nối đang có.
2. Kiểm tra hàng đợi (waiting queue) của từng gNB — nếu có thể phục vụ một UE trong queue, chuyển UE đó vào trạng thái kết nối.
3. Với mỗi UE chưa được phục vụ: chọn gNB phù hợp dựa trên chính sách (DEFAULT hoặc OFFLOAD). Nếu gNB không đủ dung lượng ngay lập tức, UE được đưa vào hàng đợi tương ứng.
4. Các kết nối hiện hành được truyền $MB - PER - STEP$ MB mỗi step; nếu UE hoàn thành gói — đánh dấu hoàn tất và giải phóng tài nguyên gNB.
5. Ghi nhận trạng thái mô phỏng (số UE đang active, số UE chờ, tải của từng gNB, v.v.) và tiến sang step tiếp theo.

3.3 Mô tả các chính sách traffic steering

3.3.1 DEFAULT

- Mỗi UE kết nối đến gNB *gần nhất* trong tập gNB có ”khả năng” — chính sách ưu gNB có chỗ trống (dựa trên tải hiện tại).
- Nếu nhiều gNB còn chỗ, chọn dựa trên khoảng cách. Nếu tất cả gNB đầy, có thể ưu gNB có nhiều chỗ trống nhất.

3.3.2 OFFLOAD

- **VOICE UEs:** ưu kết nối vào các gNB góc nếu UE nằm trong $VOICE-RADIUS$ quanh một gNB góc; nếu không, fall back sang gNB trung tâm (tùy cấu hình thực nghiệm).
- **MBB UEs:** ưu trung tâm nếu UE nằm trong $MBB - RADIUS$ quanh gNB trung tâm; nếu không, chọn gNB góc gần nhất có khả năng.
- Mục tiêu: giảm tải cho gNB trung tâm bằng cách đẩy (offload) VOICE đến các gNB góc khi khả dĩ, đồng thời tập trung MBB (lưu lượng lớn) vào trung tâm để tận dụng khả năng tập trung/điều phối.

3.4 Phân tích kỳ vọng và diễn giải kết quả

3.4.1 Throughput và tổng thời gian hoàn thành

- **OFFLOAD** có khả năng giảm tải cho gNB trung tâm nếu corner gNB có thể hấp thụ VOICE; điều này làm center nhanh chóng giải quyết MBB lớn, có thể tăng tổng throughput hệ thống.

- Tuy nhiên nếu corner có capacity quá nhỏ so với nhu cầu voice, offload sẽ tạo hàng đợi tại corners → tăng waiting time cho VOICE, có thể giảm QoS voice (nhạy cảm latency).
- **DEFAULT** có xu hướng cân bằng dựa trên khoảng cách → có thể phân phối đều hơn nếu cả 5 gNB có capacity tương tự.

3.4.2 Latency và fairness

- Nếu mục tiêu là tối đa hóa fairness (đều tải), DEFAULT có thể tốt hơn khi topologie và yêu cầu cân đối.
- Nếu mục tiêu là tối ưu QoS cho MBB (throughput cho các flows lớn), OFFLOAD ưu hoá bằng cách tập trung MBB vào center — có lợi nếu center có capacity lớn hơn.

3.4.3 Hiệu ứng bottleneck và starve

- Nếu chính sách ưu góc nhưng corners có capacity nhỏ, ta quan sát hiện tượng **bounce-back**: UE bị chờ lâu ở corners, trong khi center có thể còn khả năng phục vụ nhưng không được sử dụng (tuỳ cách implementation xử lý fallback).
- Đây là một trade-off classic giữa strict offloading (tuân theo chính sách) vs. pragmatic fallback (dựa trên tình hình thực tế).

3.5 Giới hạn mô phỏng và các giả định

- **Hệ thống chưa hoàn thiện:** Chưa train được Agent để tự động lựa chọn chính sách phù hợp, thiếu giao diện điều khiển của O-RAN.
- **Mô hình dữ liệu:** UE gửi một gói kích thước cố định và gói phải được phục vụ nguyên vẹn bởi một gNB (không phân mảnh giữa nhiều gNB).
- **Không có mobility:** UE vị trí cố định. Mobility (handovers) có thể làm thay đổi sâu sắc hiệu năng và cần được mô phỏng cho kịch bản di động.
- **Độ chính xác radio:** mô phỏng không tính đến các yếu tố vật lý phức tạp (pathloss, fading, interference, MCS adaptation). Nó giả sử coverage quyết định bằng khoảng cách đơn giản và radius ưu tiên.
- **Scheduling:** mô phỏng đơn giản hoá tài nguyên ở mức MB capacity; thực tế scheduling PHY/MAC phức tạp hơn (RB allocation, CQI, HARQ).

Chương 4

Kết luận

Báo cáo đã trình bày chi tiết các vấn đề cốt lõi liên quan đến kiến trúc mạng truy cập vô tuyến mở (Open RAN), với việc tập trung vào vai trò của các thành phần quan trọng như RIC và các giao diện mở (E2, A1, O1). Thông qua nghiên cứu, rõ ràng thấy được các lợi ích nổi bật của Open RAN so với các kiến trúc mạng truyền thống, đặc biệt là khả năng linh hoạt, tương tác đa nhà cung cấp và triển khai các ứng dụng AI và DRL thông minh trong việc quản lý tài nguyên mạng.

Phần thực nghiệm đã được thực hiện bằng việc sử dụng bộ công cụ mô phỏng RIMEDO-TS, thể hiện rõ tính ứng dụng thực tiễn của kiến trúc Open RAN và công nghệ điều khiển thông minh DRL. Kết quả thực nghiệm chứng minh rằng việc triển khai các xApp thông minh dựa trên thuật toán DQN mang lại hiệu quả cao trong việc cân bằng tải, tối ưu hóa tài nguyên và nâng cao hiệu suất tổng thể của mạng.

Kết quả nghiên cứu và thực nghiệm khẳng định rằng kiến trúc Open RAN cùng với công nghệ trí tuệ nhân tạo và học tăng cường sâu có tiềm năng lớn trong việc tái định hình và nâng cấp mạng viễn thông hiện đại. Việc tiếp tục nghiên cứu sâu hơn và mở rộng các ứng dụng này trong tương lai sẽ giúp tối ưu hóa đáng kể hiệu năng và chi phí vận hành của các nhà cung cấp dịch vụ viễn thông, góp phần quan trọng vào sự phát triển của ngành công nghiệp viễn thông nói chung.

Tài liệu tham khảo

[1] 3GPP. *Study on new radio access technology: Radio access architecture and interfaces*. 2017. URL: https://www.3gpp.org/DynaReport/38801.htm.

[2] O□RAN Software Community. *O-RAN Architecture Overview*. 2025. URL: https://docs.o-ran-sc.org/en/latest/architecture/architecture.html.

[3] Ericsson. *Open RAN: Flexibility and Interoperability in 5G Networks*. White Paper. 2023. URL: https://www.ericsson.com/en/reports-and-papers/white-papers/open-ran.

[4] A. Khurshid **andothers**. "Disaggregated RAN: O-RAN Functional Split and Deployment Insights". **in***IEEE Communications Magazine*: 62.2 (2024), **pages** 112–118.

[5] Michele Polese, Leonardo Bonati, Salvatore D'Oro, Stefano Basagni, Tommaso Melodia. "Understanding O-RAN: Architecture, Interfaces,Algorithms, Security, and Research Challenges". **in**(2022): URL: https://scispace.com/pdf/understanding-o-ran-architecture-interfaces-algorithms-2jhdy9ql.pdf#:~:text=The%20O,standardizing%20a%20virtualization%20platform%20for.

[6] Mohammad Alavirad,Umair Sajid Hashmi, Marwan Mansour, Ali Esswie, Ramy Atawia, Gwenael Poitau, Morris Repeta. "O-RAN architecture, interfaces, and standardization: Study and application to user intelligent admission control". **in**4: (2023). URL: https://www.frontiersin.org/journals/communications-and-networks/articles/10.3389/frcmn.2023.1127039/full.

[7] O-RAN Alliance. *O-RAN Architecture Description*. O-RAN WG6 White Paper. 2020. URL: https://www.o-ran.org/specifications.

[8] WG1. "O□RAN Architecture Description". **in**(2024): URL: https://www.etsi.org/deliver/etsi_ts/103900_103999/103982/08.00.00_60/ts_103982v080000p.pdf.