AWS Public Sector Blog

Mở Khóa Các Mô Hình AI Thương Mại trong AWS GovCloud (US): Truy Cập Bảo Mật Xuyên Phân Vùng với Amazon Bedrock

bởi Tyler Replogle, Doug Hairfield, Michael Pitcher, and Vin Minichino on 30 SEP 2025 in Amazon Bedrock, Artificial Intelligence, AWS GovCloud (US), Generative AI, Government, Public Sector, Technical How-to Permalink Share

Trong bài đăng này, chúng tôi sẽ trình bày ba giải pháp cho phép các workload trong AWS GovCloud (US) kết nối an toàn vào phân vùng thương mại của Amazon Web Services (AWS) để thực hiện suy luận với Amazon Bedrock. Mỗi phương pháp có những điểm mạnh, yếu riêng và đến cuối bài bạn sẽ xác định được lựa chọn phù hợp nhất cho tổ chức của mình.

Generative AI đang phát triển rất nhanh, với các foundation models (FMs) và tính năng mới được cập nhật liên tục. Những khả năng này được giới thiệu đầu tiên ở phân vùng AWS thương mại. Với những tổ chức trong môi trường nhạy cảm, chịu sự quản lý nghiêm ngặt, thách thức là làm sao bắt kịp đổi mới này và cùng lúc thử nghiệm nhanh, thu thập phản hồi khách hàng, hoàn thiện chức năng mới thành các giải pháp sẵn sàng cho nhiệm vụ thực tế.

Cách Hoạt Động của Truy Cập Xuyên Phân Vùng

Các workload trong AWS GovCloud (US) gọi Amazon Bedrock bằng cách gửi request tới API dịch vụ trong phân vùng thương mại. Thay vì gọi endpoint Amazon Bedrock trong AWS GovCloud (US), ứng dụng sẽ chuyển request qua kết nối mạng đã chọn vào phân vùng thương mại nơi Amazon Bedrock lưu trú.

Do AWS GovCloud (US) và thương mại không có kết nối nội bộ tự nhiên, hai phân vùng này được cô lập về mặt logic và vật lý. Để giao tiếp, tổ chức phải dùng đường kết nối như endpoint công khai,AWS Site-to-Site VPN hoặc AWS Direct Connect. Các kết nối này mang request giữa các phân vùng.

Khi lưu lượng truyền giữa AWS GovCloud (US) và thương mại, nó thường đi qua backbone của AWS. Như Amazon VPC FAQ giải thích: "Các gói tin xuất phát từ mạng AWS có đích đến cũng thuộc mạng AWS sẽ luôn lưu thông trên mạng toàn cầu AWS, trừ trường hợp sang khu vực Trung Quốc." Để tìm hiểu kỹ, xem bài Introduction to Network Transformation on AWS..

Amazon Bedrock sử dụng API dịch vụ qua HTTPS, mã hóa các request với TLS từ AWS GovCloud (US) sang phân vùng thương mại. Điều này đảm bảo luôn mã hóa cho giao tiếp xuyên phân vùng, bất kể cách kết nối. Đối với workload yêu cầu module mật mã đạt chuẩn FIPS 140, Bedrock cung cấp endpoint FIPS như tài liệu AWS General Reference for Amazon Bedrock.

Logic ứng dụng luôn nằm hoàn toàn trong AWS GovCloud (US).Amazon API Gateway lộ endpoint tới workload; AWS Lambda xử lý request, lấy khóa API Amazon Bedrock từ AWS Secrets Manager và gửi yêu cầu suy luận đến Bedrock ở phân vùng thương mại. Amazon Bedrock xử lý và trả về kết quả cho Lambda, sau đó trả lại ứng dụng. Amazon CloudWatch ở AWS GovCloud (US) ghi log để theo dõi hiệu suất; AWS CloudTrail ở phân vùng thương mại cung cấp nhật ký audit các API call tới Bedrock, cho biết ai gọi và lúc nào, giúp theo dõi xuyên hai phân vùng.

Nếu dùng Site-to-Site VPN hoặc AWS Direct Connect, cả hai phía GovCloud (US) và thương mại đều có Amazon Virtual Private Cloud (Amazon VPC)s nối với nhau qua kết nối này. Khác biệt duy nhất giữa endpoint công khai, VPN hay Direct Connect là đường request đi giữa hai phân vùng.

Ba Lựa Chọn Kết Nối

Có ba cách kết nối từ AWS GovCloud (US) tới phân vùng thương mại. Chọn tùy theo yêu cầu bảo mật, hiệu năng và vận hành.

Trước khi chọn, cần thiết lập xác thực bằng khóa API Amazon Bedrock ở tài khoản phân vùng thương mại Accelerate AI development with Amazon Bedrock API keys. Nên tuân

theo thực hành tốt về luân phiên khóa API này. Vì GovCloud (US) có thể chủ động kết nối ra ngoài, luân phiên khóa có thể thực hiện tự động từ GovCloud (US), giảm vận hành và tăng an ninh.

Chức năng Lambda ở GovCloud (US) lấy khóa này từ Secrets Manager và dùng xác thực với Amazon Bedrock ở phân vùng thương mại. Workload GovCloud (US) không bao giờ lưu giữ khóa trực tiếp, giảm rủi ro. Secrets Manager ở GovCloud (US) lưu trữ an toàn và chỉ Lambda truy xuất được. AWS Identity and Access Management (IAM) kiểm soát tối thiểu hóa quyền truy cập khóa.

Ngoài khóa API, cần bật quyền truy cập model trên Bedrock commercial cho từng foundation model dùng và cấu hình inference profile nếu cần như Anthropic's Claude 4.

Với VPN và Direct Connect, cần lập các endpoint riêng tư (VPC endpoint) trong VPC thương mại cho Amazon Bedrock và các dịch vụ phụ trợ như CloudWatch Logs, Secrets Manager... Endpoint này giúp lưu thông nội bộ AWS, tránh internet, kiểm soát truy cập mạnh hơn qua IAM và policy tài nguyên.

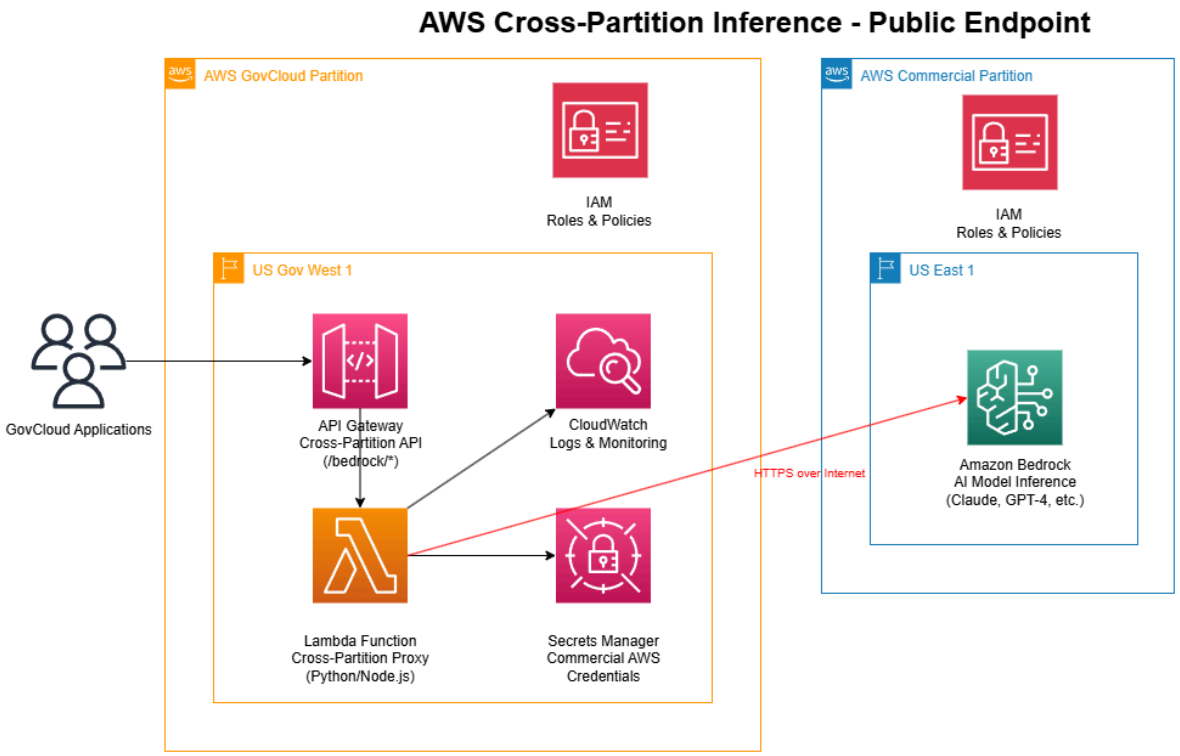
Lựa chọn 1: Endpoint công khai

Cách đơn giản nhất dùng backbone AWS để nối GovCloud (US) đến phân vùng thương mại. Ứng dụng GovCloud (US) gửi HTTPS tới API Gateway. TLS mã hóa truyền tải; IAM và Secrets Manager điều khiển xác thực, tách biệt thông tin nhạy cảm.

Yêu cầu ở phân vùng thương mại:

- Khóa API Amazon Bedrock xác thực
- Đã bật quyền truy cập model cho từng FM muốn dùng
- Inference profile cho các model yêu cầu (như Claude 4).

Giải pháp này phù hợp cho proof-of-concept hoặc dự án thử nghiệm cần triển khai nhanh, có thể xong trong vài tuần với hạ tầng tối thiểu. Đổi lại, lưu lượng đi ra internet nên không đạt mức độ bảo mật cao nhất. Sơ đồ kiến trúc minh họa dưới đây.



(Hình 1: Kiến trúc tổng quan tùy chọn endpoint công khai, ứng dụng GovCloud (US) gọi trực tiếp Bedrock ở phân vùng thương mại qua HTTPS)

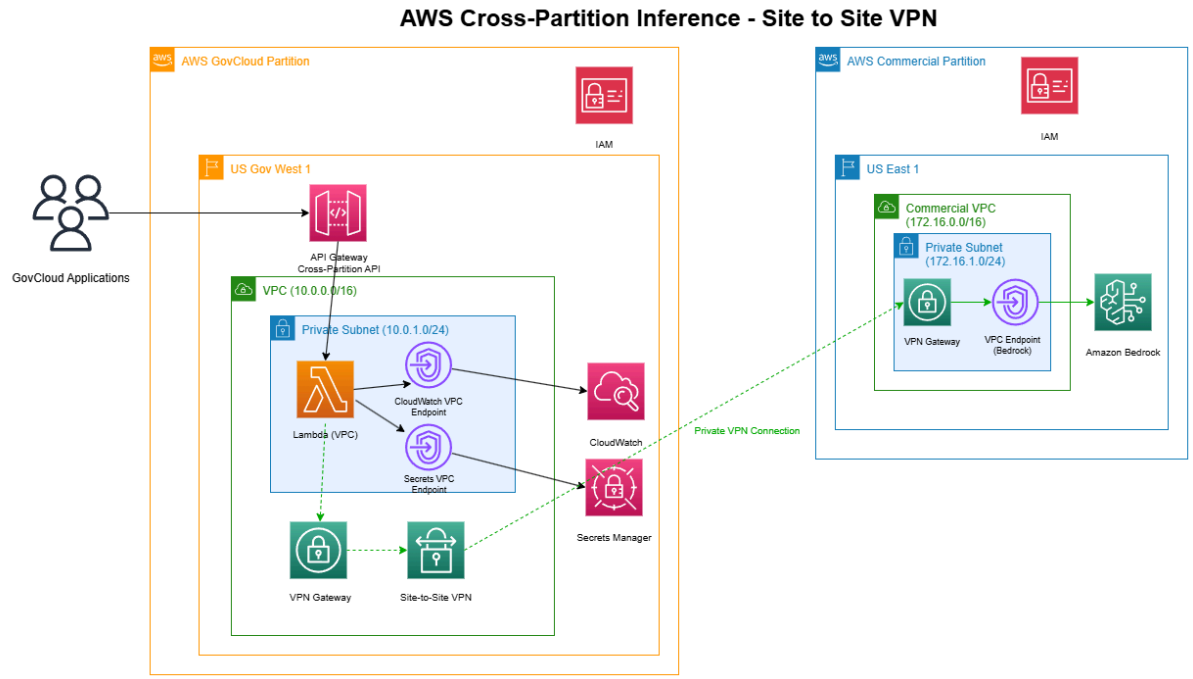
Lựa chọn 2: AWS Site-to-Site VPN với endpoint riêng tư

Dành cho tổ chức muốn bảo mật cao hơn, tránh internet công khai, dùng AWS Site-to-Site VPN lập "đường hầm" mã hóa giữa VPC GovCloud (US) và thương mại. Tất cả lưu lượng đi trong đường hầm VPN, tăng bảo mật. VPC endpoint dùng giữ lưu thông nội bộ AWS.

Yêu cầu ở phân vùng thương mại:

- Mọi thứ ở lựa chọn 1
- Đã cấu hình VPN gateway nối GovCloud (US)
- Đã tạo các VPC endpoint để dùng dịch vụ AWS mà không đi qua internet

Cách này tốt cho môi trường triển khai thực tế, tuân thủ quy định tốt hơn, bảo mật cao hơn, bù lại phức tạp hơn và tốn thời gian triển khai. Xem hướng dẫn setup chi tiết Get started with AWS Site-to-Site VPN trên AWS Docs. Sơ đồ minh họa kiến trúc dưới đây.



(Hình 2: Kiến trúc tổng quan tùy chọn kết nối VPN, GovCloud (US) và thương mại kết nối qua đường hầm VPN mã hóa)

Lựa chọn 3: AWS Direct Connect

Lựa chọn này sử dụng AWS Direct Connect tạo đường kết nối riêng tư giữa GovCloud (US) và phân vùng thương mại, cho thông lượng cao nhất, độ trễ thấp nhất, thích hợp workload AI trọng yếu hoặc khối lượng lớn. VPC endpoint giữ lưu thông trong AWS, không đi qua internet.

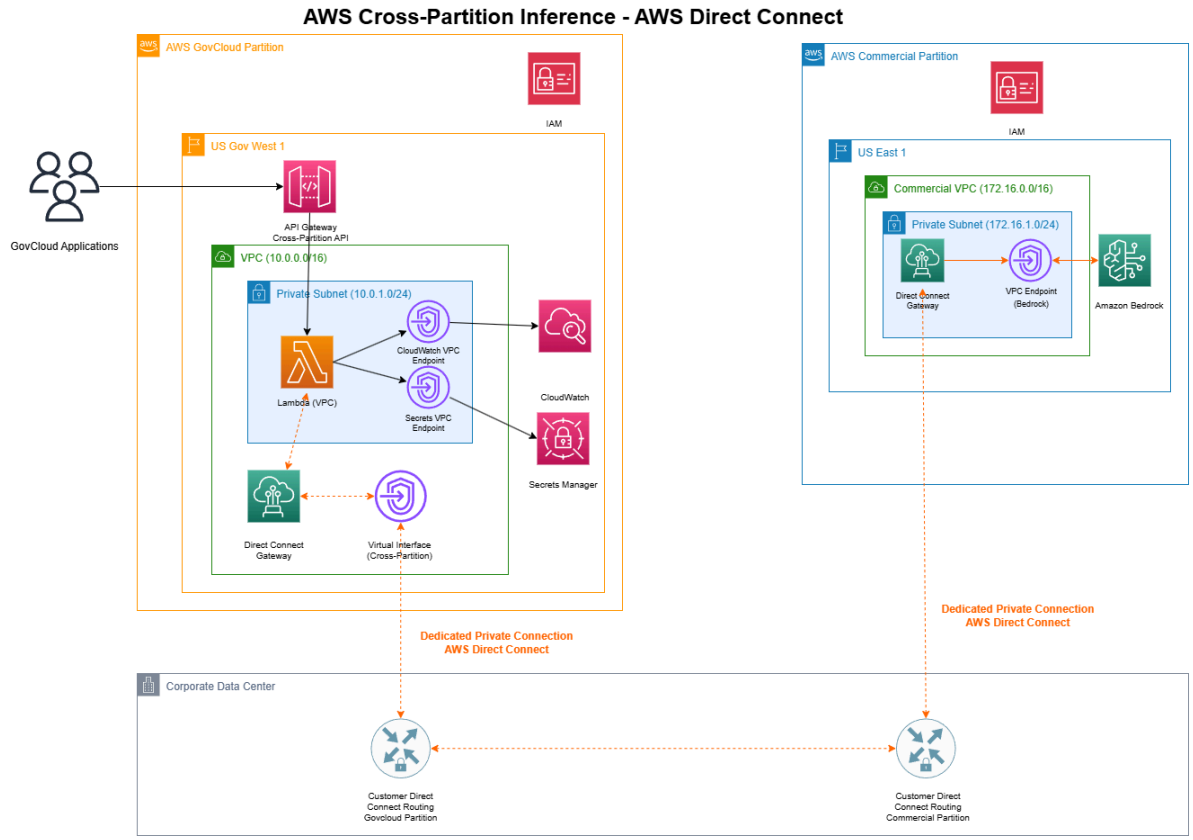
Lưu ý: Direct Connect không cung cấp đường riêng nội bộ từ GovCloud (US) sang thương mại, mà sẽ kết nối từng phân vùng tới mạng riêng của khách hàng, sau đó định tuyến giữa hai đầu ở đây. Thiết kế này cho phép kiểm soát đường đi, nhưng tăng độ trễ và phát sinh thêm cước truyền dữ liệu ở cả hai phía.

Yêu cầu ở phân vùng thương mại:

- Mọi thứ ở lựa chọn 1
- Direct Connect gateway, đa kết nối từ từng phân vùng về mạng khách hàng để định tuyến
- VPC endpoint để truy cập AWS mà không qua internet

Giải pháp này có bảo mật mạnh nhất, hiệu năng nhất quán có SLA. Tuy nhiên phải triển khai hạ tầng mạng phức tạp, tốn nhiều tuần/tháng tuỳ vào đối tác và địa điểm, thường phù hợp khi workload đã ổn định và có nhu cầu hiệu năng lớn. Xem thêm blog

về kết nối [Hybrid connectivity to AWS GovCloud (US) and commercial Regions using AWS Direct Connect](). Sơ đồ minh họa dưới đây.



(Hình 3: Kiến trúc tổng quan tùy chọn Direct Connect, từng phân vùng nối Direct Connect về mạng khách hàng, nơi định tuyến tới Bedrock ở phân vùng thương mại)

Chọn Kết Nối Phù Hợp

Chọn tùy thuộc quy định tuân thủ, độ nhạy cảm dữ liệu, hạ tầng mạng sẵn có. Với thử nghiệm, có thể dùng endpoint công khai với TLS mã hóa; còn workload thật sẽ dùng VPN hoặc Direct Connect ngay từ đầu.

Cần lưu ý: dữ liệu đi kèm prompt sẽ truyền tới dịch vụ Amazon Bedrock ở phân vùng thương mại — nghĩa là dữ liệu và ngữ cảnh đi ra khỏi GovCloud (US). Các tổ chức phải kiểm tra quy trình phù hợp quy định bảo mật, kiểm soát dữ liệu.

Cần tính chi phí truyền dữ liệu; lưu lượng đi qua phân vùng sẽ tính phí egress ở cả tài khoản GovCloud (US) lẫn thương mại. Nếu workload lớn, đây có thể là khoản tiền đáng kể.

Chọn đường kết nối không phải là bắt đầu đơn giản rồi nâng cấp, mà nên chọn khớp với chính sách bảo mật và loại dữ liệu gửi sang Bedrock ngay từ đầu. Mỗi phương án đều có mặt lợi/hại khác nhau.

Direct Connect cho phép kiểm soát đường đi tuyệt đối, dữ liệu truyền hoàn toàn qua đường riêng. Nếu đã có Direct Connect, thêm kết nối mới khá đơn giản; nếu chưa, cần thời gian nhiều tuần/tháng tùy đối tác triển khai.

Site-to-Site VPN với endpoint riêng tư là điểm trung hòa ổn (middle ground). Lưu lượng giữa GovCloud (US) và thương mại thường đi backbone AWS, VPN mã hóa truyền tải, và có thể triển khai nhanh bằng IaC. Thường dùng được với môi trường sản xuất mà không phải xây hạ tầng như Direct Connect.

Endpoint công khai dễ và nhanh nhất để triển khai, nhưng lưu lượng có thể đi ra ngoài backbone. TLS vẫn mã hóa truyền tải, song một số tổ chức sẽ không muốn dữ liệu ra khỏi mạng kiểm soát.

Kết luận

Suy luận xuyên phân vùng là giải pháp an toàn, mở rộng cho khách hàng AWS GovCloud (US) tận dụng ngay các mô hình AI mới nhất ở phân vùng thương mại. Bằng cách kết hợp mạng qua internet, VPN hoặc Direct Connect, tổ chức có thể dùng Bedrock ở phân vùng thương mại từ GovCloud (US) nếu hợp quy định và kiểm soát rủi ro.

Kiến trúc này cho phép các tổ chức vận hành trên GovCloud (US) dữ liệu nhạy cảm vẫn thử nghiệm, đổi mới với AI hiện đại ngay khi vừa ra mắt ở phân vùng thương mại. Cân bằng giữa đổi mới và kiểm soát an toàn, tuân thủ nghiêm ngặt.

TAGS: Artificial Intelligence, AWS GovCloud (US), AWS Public Sector, government, technical how-to

Tyler Replogle

Tyler là kiến trúc sư giải pháp chính và lãnh đạo về cơ sở dữ liệu kỹ thuật tại AWS cho khối công cộng toàn cầu. Anh giúp các Đối tác và khách hàng AWS vận hành các giải pháp phục vụ nhiệm vụ cuối cùng của họ trên AWS.

Doug Hairfield

Doug là kiến trúc sư giải pháp cấp cao, giúp các tổ chức khai thác sức mạnh của AI để giải quyết các vấn đề thực tế. Anh có nhiều kinh nghiệm giúp khách hàng thuộc lĩnh vực công cộng thiết kế workload trong môi trường tuân thủ cao. Khi không làm việc với giải pháp điện toán đám mây, Doug là một người cha và thích dành thời gian cho gia đình.

Michael Pitcher

Michael là quản lý cấp cao về kiến trúc giải pháp tại AWS. Trong vai trò này, anh phối hợp chặt chẽ với các đối tác để hỗ trợ mục tiêu cuối cùng của khách hàng trong khối công cộng. Michael có nhiều kinh nghiệm về bảo mật và tuân thủ, từng làm việc tại tổ chức đánh giá bên thứ ba (3PAO), nơi anh tập trung vào chứng nhận đám mây và bảo mật đám mây ở các môi trường kiểm soát nghiêm ngặt.

Vin Minichino

Vin là kiến trúc sư giải pháp cấp cao tại AWS, nơi anh hỗ trợ các đối tác dịch vụ y tế liên bang. Ngoài công việc, Vin là cha của hai con, đam mê du lịch bằng xe RV, và yêu thích chế tạo, sáng tạo.