# Unlocking Commercial AI Models in AWS GovCloud: Secure Cross-Partition Access for Government Workloads

Government agencies and regulated industries operating in AWS GovCloud face a challenge in today's AI-driven landscape: accessing the latest and most capable AI models available in the AWS Commercial partition. While AWS GovCloud provides the security and compliance features required for sensitive workloads, it has limited availability of generative AI services like Amazon Bedrock compared to the commercial partition. This creates a barrier to AI innovation for organizations that must operate within strict compliance boundaries.

In this blog We address these challenges, we've go over a a cross-partition AI inference architecture that enables GovCloud applications to securely access Amazon Bedrock services in the AWS Commercial partition. This solution maintains data security while providing access to the full range of AI models through three distinct architectural approaches, each designed for different sensitivity and performance requirements.

The solution leverages a proxy architecture that acts as a secure bridge between partitions, enabling organizations to access advanced AI capabilities without compromising their compliance posture. This approach allows government agencies to innovate with AI while maintaining the strict security controls required for their sensitive workloads.

## Benefits of Cross-Partition AI Implementation

By implementing cross-partition AI inference, government organizations can realize several transformative benefits:

**Access to Cutting-Edge AI:** Organizations gain immediate access to the latest AI models and capabilities available in AWS Commercial, enabling them to leverage state-of-the-art technology for mission-critical applications while maintaining their compliance posture.

**Maintained Compliance:** The solution ensures all data handling meets government security standards through comprehensive encryption, network isolation, and audit logging. Organizations can innovate with AI without compromising their regulatory requirements.

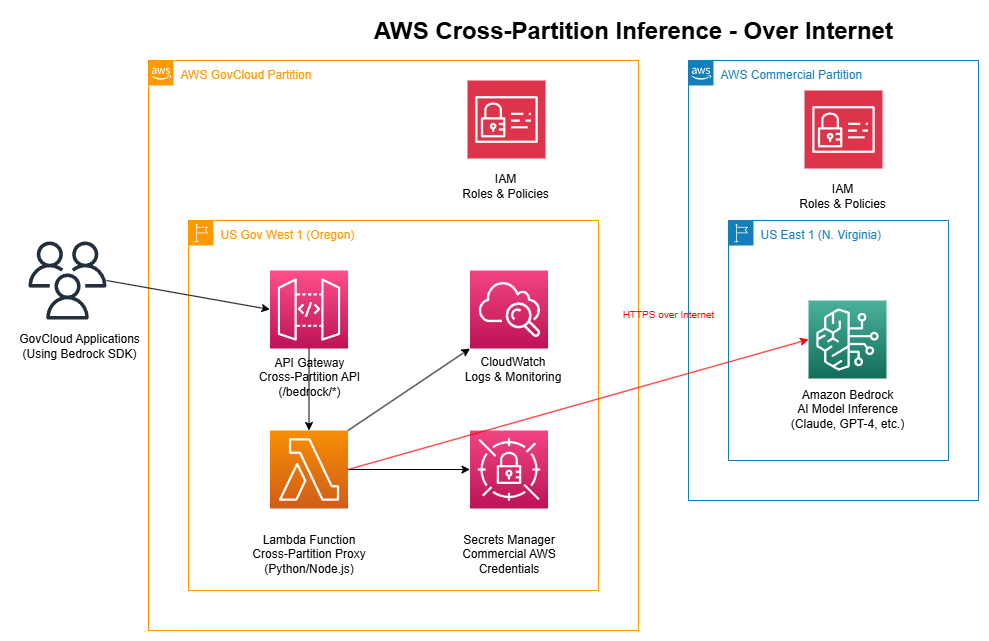
**Innovation Enablement:** The solution positions government agencies to rapidly adopt new AI capabilities as they become available, ensuring they can stay current with technological advances while meeting their security obligations.

## Three Paths to Cross-Partition AI Access

Our solution provides three implementation options, allowing organizations to choose the approach that best fits their security requirements, performance needs, and implementation timeline.

### Option 1: Internet-Based Connectivity

The Minimum Viable Product approach provides the fastest path to cross-partition AI access using HTTPS connections over the public internet, prioritizing speed of implementation while maintaining essential security controls.



\*Figure 1: Cross-Partition Inference Architecture - Over Internet (MVP)\*

This architecture demonstrates how GovCloud applications can quickly access commercial AI services through a serverless proxy. The solution includes:

1. GovCloud applications send requests to API Gateway for authentication and routing

2. Lambda function acts as a cross-partition proxy, retrieving credentials from Secrets Manager

3. HTTPS calls to Amazon Bedrock in the commercial partition over the public internet

4. Comprehensive logging through CloudWatch for audit and monitoring requirements

**Key Benefits:**

- Rapid Implementation: Can be deployed in 1-2 weeks

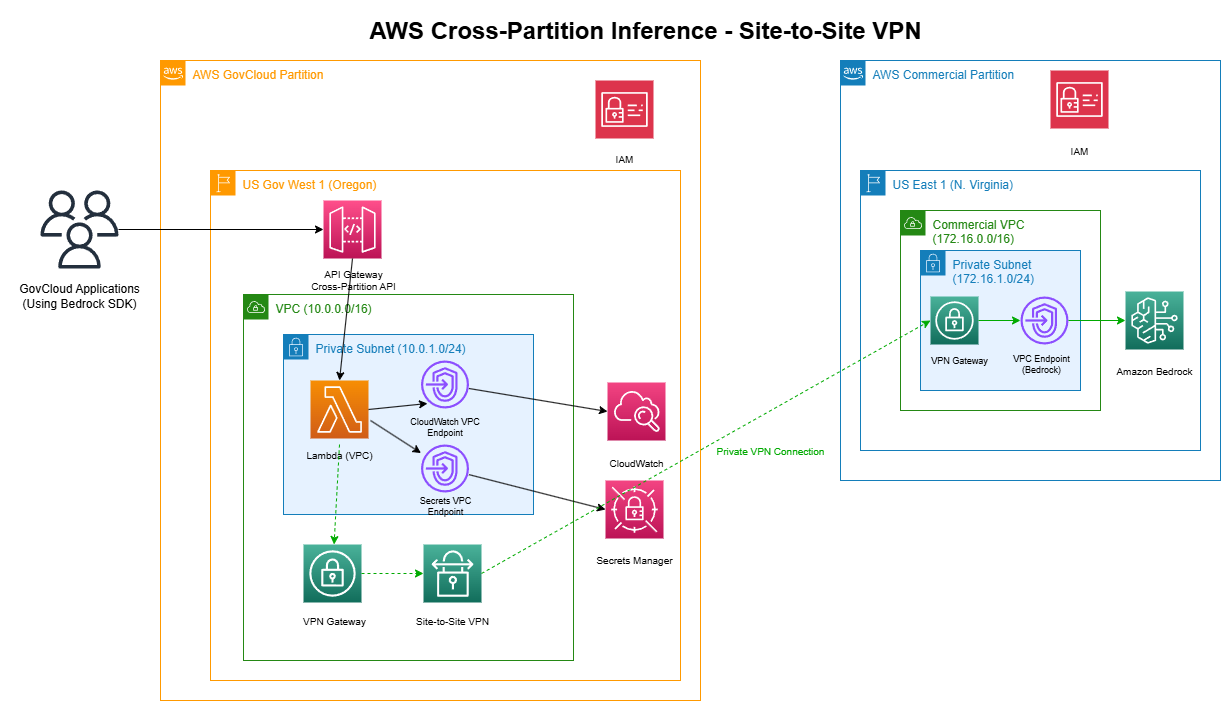
- Cost Effective: Minimal infrastructure requirements with pay-per-use model

- Simple Operations: Serverless architecture with auto-scaling capabilities

- Immediate Value: Quick access to advanced AI models for proof-of-concept work

## Option 2: Site-to-Site VPN Connectivity

The VPN approach provides a secure, encrypted tunnel between AWS partitions, ensuring all cross-partition traffic remains private while offering better performance characteristics than the internet-based solution.

\*Figure 2: Cross-Partition Inference Architecture - Site-to-Site VPN\*

This enhanced architecture features:

1.\*Private subnet deployment in both GovCloud (10.0.0.0/16) and Commercial (172.16.0.0/16) VPCs

2. VPN Gateways in each partition creating an encrypted tunnel for secure communication

3. VPC endpoints eliminating internet dependencies for AWS service access

4. Enhanced monitoring through VPC Flow Logs and comprehensive CloudWatch integration

**Key Benefits:**

- Enhanced Security: All traffic flows through private, encrypted tunnels

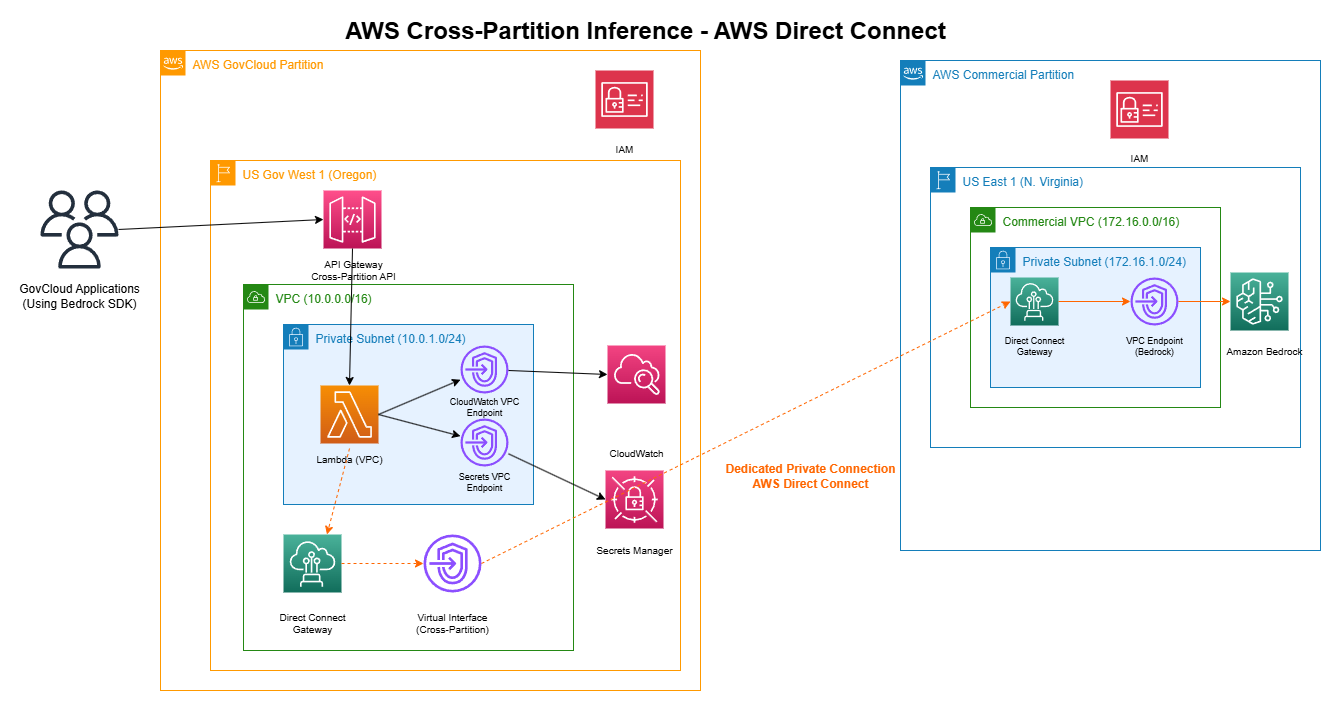
- Improved Performance: Dedicated bandwidth with reduced latency

- Better Compliance: Network isolation meets higher security requirements

- Production Ready: Suitable for production workloads with consistent performance needs

### Option 3: AWS Direct Connect

The Direct Connect solution provides the highest performance and security option through dedicated private network connections between AWS partitions, designed for high-volume, mission-critical AI workloads.

\*Figure 3: Cross-Partition Inference Architecture - AWS Direct Connect\*

This enterprise-grade architecture includes:

1. Dedicated network connections in colocation facilities with customer-managed networking equipment

2. Direct Connect Gateways in each partition providing high-bandwidth connectivity (1Gbps to 100Gbps)

3. Private connectivity with no internet transit, meeting the strictest security requirements

4. SLA-backed reliability with 99.9% uptime and redundancy options

**Key Benefits:**

- Maximum Performance: High bandwidth with consistent, low-latency connections

- Highest Security: Completely private connectivity with dedicated infrastructure

- \*\*Enterprise Scale: Supports high-volume AI inference applications

- Predictable Costs: Fixed monthly charges with lower per-GB costs at scale

## Implementation Strategy: A Phased Approach

We recommend a three-phase implementation strategy that allows organizations to start quickly while building toward enterprise-grade capabilities:

**Phase 1:** MVP Deployment (Weeks 1-4)- Establish basic cross-partition AI access using the internet-based architecture to validate functionality and gather initial performance metrics.

**Phase 2:** VPN Enhancement (Weeks 5-12)- Implement the Site-to-Site VPN architecture to improve security and performance for production workloads.

**Phase 3:** Direct Connect Optimization (Weeks 13-32) - Deploy Direct Connect infrastructure for the highest-volume, most critical applications requiring maximum performance and security.

This phased approach enables organizations to realize immediate value while building toward their long-term architectural goals, with each phase providing incremental improvements in security, performance, and capability.

### Conclusion

As government agencies worldwide seek to modernize their services and leverage AI for improved citizen outcomes, secure and scalable cross-partition AI access becomes essential infrastructure. Through our comprehensive solution combining AWS GovCloud security with commercial partition AI capabilities, organizations can build AI-powered applications that are secure, compliant, and ready for the future.

The three-tiered architecture approach ensures that organizations can start their AI journey immediately while building toward enterprise-grade capabilities that meet their most stringent requirements. This solution bridges the gap between security requirements and AI innovation, providing a clear pathway to leverage the full potential of AWS AI services.

If you'd like to learn more about implementing cross-partition AI inference solutions, discuss your specific requirements, or explore how AWS can support your AI initiatives while maintaining compliance, contact our specialist AI and government solutions team at [contact-email]. To speak with an AWS GovCloud expert and learn more about secure AI implementations, visit our [GovCloud solutions page].

\*\*TAGS:\*\* AWS GovCloud, AI/ML, Amazon Bedrock, Cross-Partition, Government, Public Sector, Best Practices, Security, Compliance