AWS Certified Machine Learning Engineer - Associate Exam Guide

# Section 3.1 — Select Deployment Infrastructure & Endpoints (In-Depth)

This section focuses on selecting the right deployment infrastructure and endpoints for machine learning models. It covers deployment types, compute resources, orchestration options, and AWS services that support production ML workloads.

## 1. Deployment Types

• Real-time inference → low-latency APIs (fraud detection, chatbots).

• Batch inference → scheduled large-scale jobs (recommendations, daily reports).

• Asynchronous inference → handle long-running tasks without client timeout (e.g., document processing).

• Edge deployment → use SageMaker Neo to optimize models for edge devices (IoT, mobile).

## 2. Compute Resources

• CPU instances → cheaper, good for lightweight inference.

• GPU instances → required for deep learning, image/video, NLP models.

• Memory-optimized instances → large datasets or models requiring high throughput.

• Provisioned vs On-Demand → choose based on workload predictability.

## 3. Endpoint Options

• Single-model endpoint → host one model per endpoint.

• Multi-model endpoint → dynamically load multiple models on demand (cost-efficient).

• Multi-container endpoint → host different containers for diverse models.

• Serverless inference → scales automatically with usage, cost-efficient for sporadic requests.

• Async inference → decouples client requests from model execution for long jobs.

## 4. Orchestration & Targeting

• SageMaker hosting services → manage real-time, batch, and async endpoints.

• SageMaker Pipelines → automate training-to-deployment workflows.

• Amazon ECS/EKS → container orchestration for ML deployment.

• AWS Lambda → lightweight inference layer using SageMaker endpoints.

• Shadow/Canary deployments → test new versions safely.

## 5. AWS Tools

• SageMaker Endpoints → production hosting for models.

• SageMaker Neo → compile and optimize models for edge.

• Elastic Inference → attach GPU acceleration to CPU instances for cost savings.

• Amazon API Gateway + Lambda → expose ML models as REST APIs.

• AWS Step Functions → orchestrate inference pipelines.

## ✅ Exam Tips

• Real-time = latency sensitive, Batch = periodic, Async = long jobs.

• Multi-model endpoints = cost optimization when hosting many models.

• SageMaker Neo = edge deployments, Elastic Inference = GPU cost savings.

• Serverless = sporadic workloads, pay-per-request.

• Canary/Shadow deployments = safe rollout strategies.