AWS Certified Machine Learning Engineer - Associate Exam Guide

# Section 4.2 — Monitor & Optimize Infrastructure and Costs (In-Depth)

This section focuses on monitoring ML infrastructure performance, optimizing resource utilization, and managing costs. It includes observability tools, cost tracking, scaling practices, and AWS services for optimization.

## 1. Infrastructure KPIs

• Utilization → CPU, GPU, memory usage.

• Throughput → requests/second or data processed per batch job.

• Latency → response time per request.

• Availability & Fault Tolerance → uptime, error handling, failover capacity.

• Scalability → ability to handle workload increases.

## 2. Observability & Monitoring Tools

• Amazon CloudWatch → monitor CPU/GPU/memory, logs, create alarms/dashboards.

• AWS CloudTrail → track API calls, auditing, triggering retraining pipelines.

• AWS X-Ray → trace distributed applications for bottlenecks.

• Lambda Insights & CloudWatch Logs Insights → analyze serverless workloads.

• QuickSight → visualize monitoring metrics.

## 3. Cost Optimization

• AWS Cost Explorer → analyze usage patterns, forecast costs.

• AWS Budgets → set cost and usage budgets with alerts.

• AWS Trusted Advisor → recommends cost savings and best practices.

• Tagging resources → track costs by project/team.

• Instance purchasing options: Spot (cheapest), On-Demand (flexible), Reserved (predictable), Savings Plans.

## 4. Scaling & Right-Sizing

• Auto Scaling → dynamically adjust instance counts.

• SageMaker Inference Recommender → suggests best instance types for inference workloads.

• AWS Compute Optimizer → recommends right-sized instances based on usage.

• Provisioned Concurrency for Lambda → balance latency vs cost.

• Batch Transform → cost-efficient for large batch inference jobs.

## 5. Best Practices

• Always monitor GPU utilization to avoid under/overprovisioning.

• Use Spot Instances for cost savings in training jobs.

• Right-size instances regularly using Compute Optimizer.

• Use tagging and cost allocation reports for accountability.

• Configure CloudWatch alarms to detect anomalies early.

## ✅ Exam Tips

• Inference Recommender = find best instance type for production endpoints.

• Compute Optimizer = right-sizing instances for cost/performance.

• Spot = cheapest for training, Reserved = best for long-term predictable workloads.

• CloudWatch = logs/metrics, CloudTrail = API auditing.

• Cost Explorer + Budgets = track and manage AWS spend.