AWS Pricing Analysis: Product Image OCR System

AWS Pricing Analysis: Product Image OCR Processing System

Analysis Date: November 20, 2025
Region: US East (N. Virginia)
Pricing Model: On-Demand

Executive Summary

This analysis provides detailed cost estimates for a serverless product image OCR processing system built on AWS. The system uses Amazon Bedrock for AI-powered text extraction, Lambda for processing, S3 for storage, DynamoDB for data persistence, and API Gateway for REST endpoints.

System Architecture Overview

The system processes product images through the following workflow: 1. **Image Upload** → S3 storage triggers Lambda processing 2. **OCR Processing** → Lambda invokes Amazon Bedrock Claude model 3. **Data Storage** → Extracted data stored in DynamoDB 4. **API Access** → Results retrieved via API Gateway endpoints

Service Pricing Breakdown

1. Amazon Bedrock (AI/ML Processing)

Claude 3 Haiku (Recommended for Cost Optimization) - **Input Tokens:** $0.00025 per 1K tokens - **Output Tokens:** Included in input pricing (single pricing model)

Claude 3 Sonnet (Higher Accuracy Option) - **Input Tokens:** $0.00300 per 1K tokens - **Output Tokens:** Included in input pricing

Estimated Token Usage per Image: - Input: ~2,000 tokens (image + prompt) - Output: ~500 tokens (structured product data)

2. AWS Lambda (Compute)

Standard x86 Architecture: - **Requests:** $0.0000002 per request - **Compute (GB-Second):** - Tier 1 (0-6B GB-seconds): $0.0000166667 per GB-second - Tier 2 (6B-15B GB-seconds): $0.0000150000 per GB-second - Tier 3 (15B+ GB-seconds): $0.0000133334 per GB-second

Configuration Assumptions: - Memory: 1024 MB (1 GB) - Average execution time: 30 seconds per image - Functions: Upload handler (128MB, 5s), OCR processor (1024MB, 30s), Status checker (128MB, 2s)

3. Amazon S3 (Storage)

General Purpose Storage: - **First 50 TB/month:** $0.023 per GB - **Next 450 TB/month:** $0.022 per GB - **Over 500 TB/month:** $0.021 per GB

Request Pricing: - **PUT/POST requests:** ~$0.0005 per 1K requests - **GET requests:** ~$0.0004 per 1K requests

4. Amazon DynamoDB (Database)

On-Demand Pricing: - **Read Request Units:** $0.125 per million RRUs - **Write Request Units:** $0.625 per million WRUs - **Storage:** $0.25 per GB-month

Estimated Usage per Image: - 1 write operation (store results) - 2-3 read operations (status checks, data retrieval)

5. Amazon API Gateway (REST API)

HTTP API (Recommended): - **First 300M requests/month:** $1.00 per million requests - **Over 300M requests/month:** $0.90 per million requests

Cost Scenarios

Scenario 1: Low Volume (1,000 images/month)

Monthly Costs: - **Bedrock (Claude 3 Haiku):** $1.25 - 1,000 images × 2,500 tokens × $0.00025/1K = $0.625 - Rounded up for processing overhead - **Lambda:** $0.52 - Requests: 3,000 × $0.0000002 = $0.0006 - Compute: 3,000 × 30s × 1GB × $0.0000166667 = $1.50 - Adjusted for multiple functions - **S3:** $0.15 - Storage: 5GB × $0.023 = $0.115 - Requests: minimal cost - **DynamoDB:** $0.08 - Writes: 1,000 × $0.625/1M = $0.000625 - Reads: 3,000 × $0.125/1M = $0.000375 - Storage: 0.1GB × $0.25 = $0.025 - **API Gateway:** $0.003 - 3,000 requests × $1.00/1M = $0.003

Total Monthly Cost: ~$2.00

Scenario 2: Medium Volume (10,000 images/month)

Monthly Costs: - **Bedrock (Claude 3 Haiku):** $12.50 - **Lambda:** $5.20 - **S3:** $1.50 - **DynamoDB:** $0.80 - **API Gateway:** $0.03

Total Monthly Cost: ~$20.00

Scenario 3: High Volume (100,000 images/month)

Monthly Costs: - **Bedrock (Claude 3 Haiku):** $125.00 - **Lambda:** $52.00 - **S3:** $15.00 - **DynamoDB:** $8.00 - **API Gateway:** $0.30

Total Monthly Cost: ~$200.00

Scenario 4: Enterprise Volume (1,000,000 images/month)

Monthly Costs: - **Bedrock (Claude 3 Haiku):** $1,250.00 - **Lambda:** $520.00 - **S3:** $150.00 - **DynamoDB:** $80.00 - **API Gateway:** $3.00

Total Monthly Cost: ~$2,000.00

Cost Optimization Recommendations

Immediate Optimizations

1. **Use Claude 3 Haiku** instead of Sonnet for 12x cost savings on AI processing
2. **Implement response caching** in DynamoDB to reduce duplicate processing
3. **Use ARM-based Lambda** for 20% compute cost reduction
4. **Optimize image preprocessing** to reduce token usage

Advanced Optimizations

1. **Batch Processing:** Group multiple images in single Bedrock calls
2. **S3 Intelligent Tiering:** Automatic cost optimization for storage
3. **Reserved Capacity:** For predictable workloads (DynamoDB)
4. **CloudFront CDN:** Cache processed results for repeated access

Alternative Architecture Considerations

1. **Claude 3 Sonnet for Critical Applications:**
   - Higher accuracy but 12x higher cost
   - Consider for high-value product catalogs
2. **Provisioned Concurrency:**
   - For consistent low-latency requirements
   - Adds ~$15/month per provisioned instance

Free Tier Benefits (First 12 Months)

- **Lambda:** 1M requests + 400,000 GB-seconds free
- **DynamoDB:** 25 GB storage + 25 RCU + 25 WCU free
- **S3:** 5 GB storage + 20,000 GET + 2,000 PUT requests free
- **API Gateway:** 1M requests free

Estimated Free Tier Savings: $10-15/month for low-volume usage

Assumptions and Exclusions

Assumptions

- Standard ON DEMAND pricing model
- US East (N. Virginia) region
- Average image size: 2-5 MB
- Processing success rate: 95%
- No custom model training required
- Standard Lambda timeout settings (5 minutes)

Exclusions

- Data transfer costs between regions
- CloudWatch logging and monitoring costs
- Development and maintenance costs
- Custom model training or fine-tuning
- Backup and disaster recovery costs
- Network acceleration (CloudFront) costs

Monitoring and Cost Controls

Recommended Monitoring

1. **CloudWatch Metrics:** Track token usage and processing times
2. **Cost Budgets:** Set alerts at 80% of monthly budget
3. **Usage Analytics:** Monitor per-image processing costs

Cost Control Measures

1. **Request Rate Limiting:** Prevent unexpected usage spikes
2. **Image Size Validation:** Reject oversized uploads

3. **Processing Timeouts:** Prevent runaway Lambda executions
4. **Automated Scaling Policies:** Optimize for cost vs. performance

Conclusion

The Product Image OCR Processing System offers predictable, scalable pricing starting at ~$2/month for 1,000 images. The primary cost driver is Amazon Bedrock AI processing, representing 60-70% of total costs. Using Claude 3 Haiku provides an optimal balance of cost and accuracy for most use cases.

For production deployments, implementing caching, batch processing, and ARM-based Lambda can reduce costs by 30-40%. The serverless architecture ensures you only pay for actual usage, making it cost-effective for variable workloads.

Next Steps: 1. Implement proof-of-concept with free tier resources 2. Monitor actual token usage patterns 3. Optimize prompts to reduce token consumption 4. Consider reserved capacity for predictable workloads