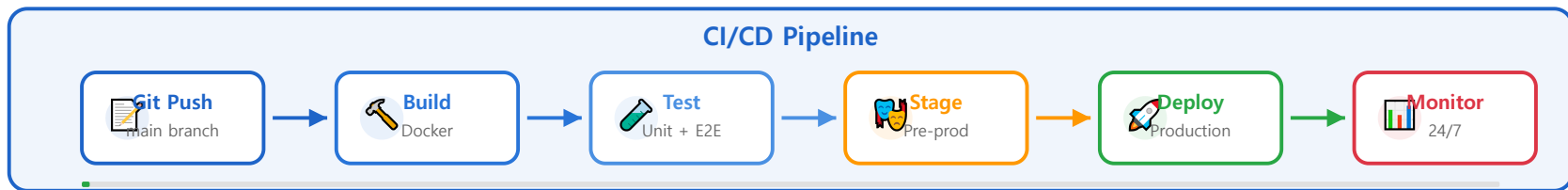
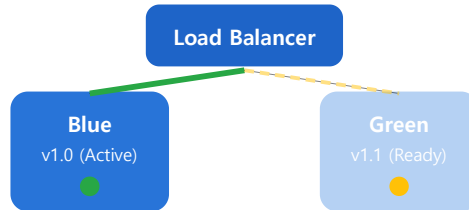


# Deployment Strategies

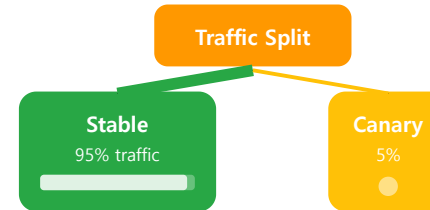
## RAG System Deployment Architecture



### Blue-Green Deployment



### Canary Deployment



### Real-time Monitoring

P95 Latency: **82ms** ✓

Error Rate: **0.05%** ✓

QPS: **1,200** ✓

CPU: **68%**

**82ms**

P95 Latency

**0.05%**

Error Rate

**1,200**

QPS

**68%**

CPU Usage