

Model Serving Architecture Patterns

Dedicated Servers

One model per server

Predictable latency
Isolated resources

Latency: 10-50ms
Cost: High

Multi-Model Serving

Multiple models shared

Better utilization
Lower cost per model

Latency: 20-100ms
Cost: Medium

Serverless

Auto-scaling functions

Pay per use
Cold start latency

Latency: 100ms-5s
Cost: Low (variable)

Selection Guide

Dedicated Servers:

- Strict latency requirements (less than 50ms p99)
- High request volume
- Predictable performance critical

Multi-Model Serving:

- Multiple models with moderate traffic
- Cost optimization priority
- Relaxed latency (100-200ms acceptable)

Serverless:

- Unpredictable or bursty traffic
- Cold start latency acceptable
- Batch or async processing