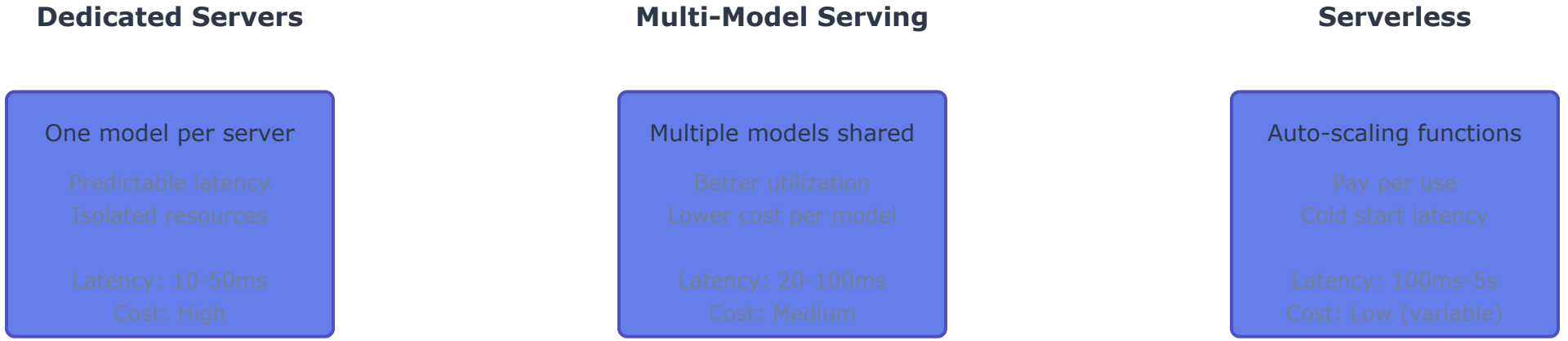
Model Serving Architecture Patterns



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