# The 50ms Standard

## Engineering Real-Time Financial Intelligence

● AI/ML Technical Documentation ● Production Systems ● Fintech Innovation ●

Technical Documentation Series

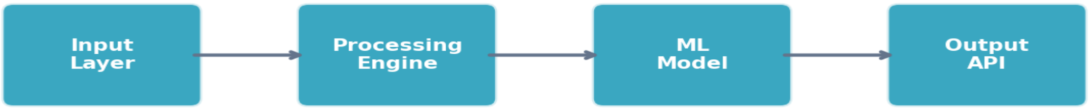
December 2024

## Executive Summary

Sub-50ms latency is the new standard for financial intelligence systems. This article explores the engineering challenges and architectural patterns required to achieve real-time financial data processing at scale.

| | | |
|---|---|---|
| **<50ms**<br>Target Latency | **20M+**<br>Requests/Day | **99.99%**<br>Uptime SLA |
| **15ms**<br>P50 Latency | **47ms**<br>P99 Latency | **8.2K**<br>Peak RPS |

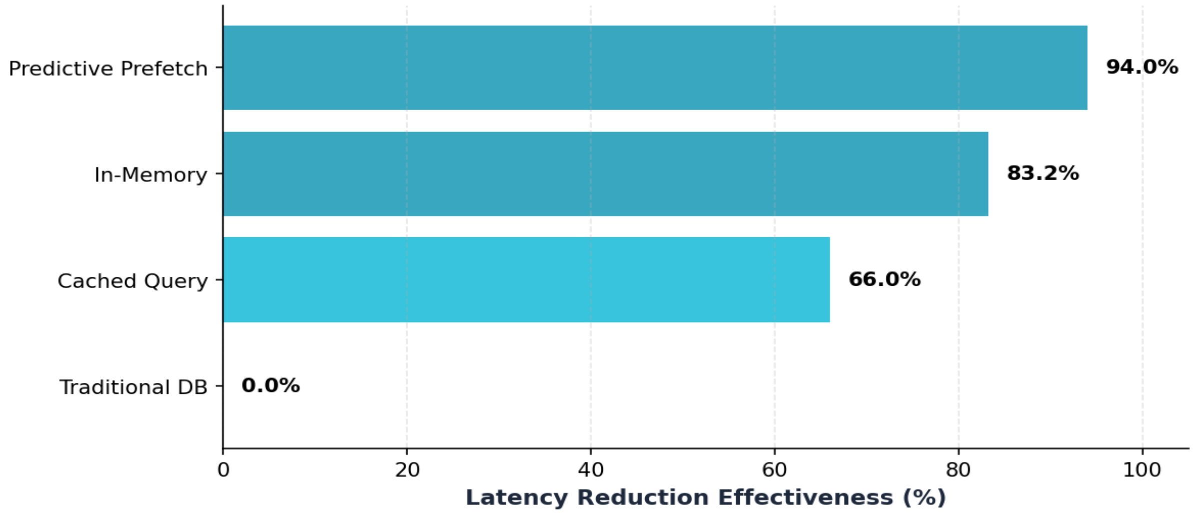
## Architecture for Speed



### • Key Components

- **Stream Processing:** Kafka + Apache Flink for real-time data pipelines
- **In-Memory Databases:** Redis for sub-millisecond data access
- **Edge Computing:** Distributed nodes near data sources
- **HTTP/2 Multiplexing:** Parallel request handling

## Performance Optimization Techniques



**Plaid's Achievement:** Processing 20M+ daily requests with P99 latency under 50ms through aggressive caching, predictive prefetching, and edge deployment.

## Implementation Strategies

### • 1. Streaming Architecture

- Event-driven design with Kafka topics
- Apache Flink for stateful stream processing
- Exactly-once semantics for financial accuracy

### • 2. Caching Strategy

- Multi-layer cache (CDN → Redis → Database)
- Intelligent cache warming based on usage patterns
- Cache invalidation strategies for consistency

### • 3. Monitoring & Observability

- Real-time latency tracking per endpoint
- Distributed tracing for bottleneck identification
- Automated alerting on SLA breaches