

Real-time BSS: Latency Benchmarks

Performance benchmarks for real-time rating and charging engines under a 100M+ subscriber load with sub-5ms response targets.

18 Min Read · TB-2023-088 · November 2023

EXECUTIVE SUMMARY

Real-time charging at Tier-1 scale demands consistent sub-5ms response under peak load without sacrificing rating accuracy. This brief presents benchmark results from a 100 million-subscriber synthetic load test across three charging engine configurations, and the architectural decisions that most affected tail latency.

4.2ms

P99 response time

100M+

Simulated subscribers

0.0002%

Rating error rate



Benchmark methodology

Testing simulated a realistic subscriber mix across prepaid, postpaid, and IoT rate plans at Tier-1 scale, with load patterns modeled on observed peak-hour traffic from operator deployments. Each configuration was tested across sustained load, burst load, and degraded-dependency scenarios.

- Sustained load: 100M subscribers, steady-state transaction rate
- Burst load: 3x traffic spike simulating promotional or emergency events
- Degraded dependency: charging engine response with a delayed upstream rating source



Results: response time and accuracy under load

The in-memory rating cache configuration consistently outperformed the database-backed and hybrid configurations on tail latency, at the cost of a more complex cache-invalidation model during rate-plan changes.

HEADLINE RESULT

P99 response time held at 4.2ms under sustained 100M-subscriber load, with rating error rate of 0.0002% — within target tolerance for real-time billing integrity.

Where latency was lost

The majority of tail-latency outliers traced to cache invalidation during concurrent rate-plan updates, rather than to the core rating computation itself — pointing to invalidation strategy as the highest-leverage optimization target.



Architectural recommendations

For operators targeting sub-5ms charging response at Tier-1 scale, we recommend an in-memory rating cache with write-through invalidation, paired with an asynchronous reconciliation pass against the source-of-truth rating store.

Continue the conversation.

To discuss how the findings in “Real-time BSS: Latency Benchmarks” apply to your network, request a full institutional briefing with our solutions architecture team.

[Request a Demo](#)

[Visit zoikonex.com](https://zoikonex.com)

© 2026 ZoikoNex · ZoikoNex is an AI-powered telecom monetization, OSS/BSS and governed billing platform operated by Zoiko Tech Inc., the technology arm of Zoiko Group Inc.