■ Ambient API Performance Test Report

Comprehensive Load Testing Analysis - 30 Concurrent Users

API Endpoint	https://innovationz-qa.myqone.com/Ambient/generate_summary_html_v1
Test Method	POST Request Load Testing
Concurrent Users	30 Users
Test Duration	89.4 seconds
Testing Tool	Locust Framework
Test Start Time	10:08:14
Test End Time	10:09:44
Total Requests	102
Success Rate	100.0%
Error Rate	0.0%
Report Generated	2025-07-25 10:09:44

■ Executive Summary

This performance test was conducted on July 25, 2025, with 30 concurrent users over 89.4 seconds. The system processed 102 requests with a 100.0% success rate and 0.0% error rate. While system reliability remained excellent, the average response time of 20.19 seconds indicates significant performance challenges that require attention before production deployment.

Performance Test Results

Metric	Value	Status	Benchmark
Total Requests	102	■	N/A
Successful Requests	102 (100.0%)	■ EXCELLENT	>95%
Failed Requests	0 (0.0%)	■ EXCELLENT	<5%
Average Response Time	20,190 ms	■ CRITICAL	<2000ms
Median Response Time	21,876 ms	■ CRITICAL	<1500ms
Min Response Time	3,861 ms	■ SLOW	<500ms
Max Response Time	24,824 ms	■ CRITICAL	<10000ms
95th Percentile	23,387 ms	■ CRITICAL	<3000ms
99th Percentile	24,477 ms	■ CRITICAL	<5000ms
Throughput	1.14 req/sec	■ LOW	>10 req/sec
Error Rate	0.0%	■ EXCELLENT	<1%

System Resource Utilization

Resource	Average	Maximum	Status
CPU Usage	43.5%	100.0%	■■ NORMAL-HIGH
Memory Usage	85.5%	86.7%	■ HIGH

Analysis:
• **CPU Utilization:** Moderate at 43.5% average but peak at 100.0%, indicating CPU bottlenecks during processing
• **Memory Utilization:** High at 85.5% average, showing significant memory pressure
• **Test Duration:** 89.4 seconds for 102 requests shows poor throughput at 1.14 req/sec
• **Performance Characteristics:** Consistent high response times across all requests

Detailed Performance Analysis

Response Time Distribution

The test revealed consistent performance issues across all 102 requests:

1. High Average Response Time: 20.19 seconds
Significantly exceeds acceptable limits for web API responses. Target should be under 2 seconds.

2. Consistent High Latency Pattern
• Minimum: 3.86s (still above ideal thresholds)
• Median: 21.88s (shows systemic performance issues)
• Maximum: 24.82s (approaching timeout limits)

3. Excellent Reliability: 100.0% Success Rate
• All 102 requests completed successfully
• Zero errors or timeouts (0.0% error rate)
• System maintains stability despite performance challenges

4. Low Throughput: 1.14 requests/second

• Well below industry standards (target: >10 req/sec)
• Indicates processing bottlenecks
• Limits system scalability potential

5. Resource Utilization Concerns:
• High memory usage: 85.5% average
• CPU spikes to 100.0% during processing
• Resource inefficiency relative to output

Percentile Analysis

Percentile	Response Time (ms)	Response Time (seconds)	Assessment
50th (Median)	21,876	21.88s	■ Critical
95th	23,387	23.39s	■ Critical
99th	24,477	24.48s	■ Critical
Min	3,861	3.86s	■■ Slow
Max	24,824	24.82s	■ Critical

Performance Issues Identified

■ KEY PERFORMANCE ISSUES - ANALYSIS RESULTS

1. Unacceptable Response Times
• Average: 20.19 seconds (target: <2 seconds)
• Median: 21.88 seconds (target: <1.5 seconds)
• Maximum: 24.82 seconds (approaching timeout thresholds)
• **Impact:** Users will experience unacceptable delays, likely leading to abandonment

2. Extremely Low Throughput
• Current: 1.14 requests/second
• Industry standard: >10 requests/second
• Processed only 102 requests in 89.4 seconds
• **Impact:** System cannot handle meaningful production load

3. High Resource Utilization with Poor Output
• Memory usage: 85.5% average (concerning levels)
• CPU spikes: Up to 100.0% during processing
• Resource efficiency: Very poor given low throughput
• **Impact:** Indicates architectural inefficiencies

4. Positive: Perfect Reliability
• 100.0% success rate across all requests
• 0.0% error rate (excellent stability)
• No timeouts or failures during 89.4 seconds test
• **Impact:** System is stable but slow

5. Scalability Concerns
• 30 concurrent users already showing severe performance degradation
• Response times consistently high across all percentiles
• No indication of performance improvement with optimization
• **Impact:** Production deployment with current performance impossible

Test Summary

TEST EXECUTION SUMMARY:
• Test Date: July 25, 2025
• Test Duration: 89.4 seconds (10:08:14 - 10:09:44)
• Concurrent Users: 30
• Total Requests: 102
• Success Rate: 100.0% (All requests successful)
• Average Response Time: 20.19 seconds
• Throughput: 1.14 requests/second
• **VERDICT: PERFORMANCE OPTIMIZATION REQUIRED**

Performance Optimization Recommendations

Immediate Actions (Critical Priority)

1. Performance Optimization Initiative
• Target 90% response time reduction (from 20.2s to <2s)
• Implement aggressive caching strategies
• Optimize AI/ML model processing pipeline
• Review and optimize database queries

2. Architecture Review
• Evaluate current synchronous processing model
• Consider asynchronous processing for heavy operations
• Implement connection pooling and resource management
• Review memory usage patterns and optimization

3. Capacity Planning
• Current safe capacity: 30 users with 20.2s response times
• Target capacity: 30+ users with <2s response times
• Establish performance monitoring and alerting

Short-term Improvements

1. System Monitoring
• Implement real-time performance monitoring
• Set up alerts for response times >5 seconds
• Monitor resource utilization trends
• Track throughput and error rates

2. Performance Testing Protocol
• Establish baseline performance metrics
• Regular regression testing after optimizations
• Load testing with incremental user counts
• Performance profiling to identify bottlenecks

3. Resource Optimization
• Memory usage optimization (currently 85.5%)
• CPU spike investigation and mitigation
• I/O and network optimization
• Code profiling and optimization

Performance vs Industry Standards

Metric	Current Performance	Industry Standard	Gap Analysis
Avg Response Time	20.19s	<2s	Exceeds by 18.2s
95th Percentile	23.39s	<3s	Exceeds by 20.4s
Throughput	1.14 req/s	>10 req/s	Below by 8.9 req/s
Success Rate	100.0%	>95%	■ Exceeds
Error Rate	0.0%	<1%	■ Meets
Memory Usage	85.5%	<80%	Exceeds by 5.5%

Test Conclusion & Next Steps

The performance test conducted on **July 25, 2025** with **30 concurrent users** reveals significant performance challenges:

■ POSITIVE FINDINGS:
- ■ **Perfect Reliability:** 100.0% success rate across 102 requests
- ■ **System Stability:** 0.0% error rate shows robust error handling
- ■ **Consistent Performance:** Response times are predictable (though high)
- ■ **No Failures:** All requests completed successfully within 89.4 seconds

■ CRITICAL PERFORMANCE ISSUES:
- ■ **Response Time:** 20.19s average (10x slower than target)
- ■ **Throughput:** 1.14 req/sec (far below industry standards)
- ■ **Resource Efficiency:** High memory usage (85.5%) with poor output
- ■ **Scalability:** Current architecture insufficient for production load

■ BUSINESS IMPACT ASSESSMENT:
The current performance characteristics would result in:
- User frustration due to 20.2-second wait times
- Inability to serve meaningful concurrent user loads
- Competitive disadvantage due to poor user experience
- High infrastructure costs relative to throughput

■ RECOMMENDED ACTION PLAN:
1. **IMMEDIATE:** Performance optimization initiative (target: <2s response time)
2. **SHORT-TERM:** Architecture review and optimization
3. **ONGOING:** Performance monitoring and regular testing
4. **VALIDATION:** Re-test after optimization to validate improvements

■ SUCCESS CRITERIA FOR NEXT TEST:
- Average response time: <2 seconds
- Throughput: >10 requests/second
- Memory usage: <80%
- Maintain: 100.0% success rate

■ VERDICT: OPTIMIZATION REQUIRED BEFORE PRODUCTION
While the system demonstrates excellent reliability, performance optimization is essential for production readiness.

Report Generated: 2025-07-25 14:32:51
Source Data: ambient_api_performance_report_30users_20250725_100944.html
Test Execution: 2025-07-25 10:09:44 (10:08:14 - 10:09:44)
Analysis by: Automated Performance Testing Framework
STATUS: PERFORMANCE OPTIMIZATION REQUIRED