■ Ambient API Performance Test Report

Heavy Load Testing Analysis - 40 Concurrent Users

API Endpoint	https://innovationz-qa.myqone.com/Ambient/generate_summary_html_v1
Test Method	POST Request Load Testing
Concurrent Users	40 Users (Heavy Load)
Test Duration	78 seconds
Testing Tool	Locust Framework
Test Start Time	12:37:00
Test End Time	12:38:18
Total Requests	125
Success Rate	96.8%
Error Rate	3.2%
Report Generated	2025-07-25 12:38:18

■ Executive Summary

This comprehensive performance test was conducted on July 25, 2025, at 12:38:18 with 40 concurrent users over 78 seconds. The system processed 125 requests with a 96.8% success rate and 3.2% error rate. The average response time of 18.5 seconds indicates performance challenges under heavy load, while the system demonstrates reasonable reliability with some errors appearing under stress conditions.

Performance Test Results

Metric	Value	Status	Target
Total Requests	125	■ Good	N/A
Successful Requests	121 (96.8%)	■ Good	>95%
Failed Requests	4 (3.2%)	■■ Acceptable	<5%
Average Response Time	18,500 ms	■ Poor	<2000ms
Median Response Time	19,200 ms	■ Poor	<1500ms
Min Response Time	4,200 ms	■■ Slow	<500ms
Max Response Time	35,800 ms	■ Very Slow	<10000ms
95th Percentile	31,000 ms	■ Poor	<3000ms
99th Percentile	34,500 ms	■ Poor	<5000ms
Throughput	1.6 req/sec	■■ Low	>5 req/sec
Error Rate	3.2%	■■ Acceptable	<1%

System Resource Utilization

Resource	Average	Maximum	Status
CPU Usage	52.3%	100.0%	■■ High
Memory Usage	78.2%	85.4%	■■ High

Analysis:
• **CPU Utilization:** High at 52.3% average with peaks at 100.0%, indicating significant processing bottlenecks
• **Memory Utilization:** High at 78.2% average, showing substantial memory pressure under load
• **Test Duration:** 78 seconds for 125 requests shows throughput of 1.6 req/sec
• **Resource Efficiency:** High resource usage with moderate throughput indicates potential optimization needs
• **Error Emergence:** 3.2% error rate suggests system approaching capacity limits

Detailed Performance Analysis

Response Time Analysis

The test with 40 concurrent users reveals important insights about system behavior under heavy load:

1. Heavy Load Performance: 18.5 seconds average
System shows degraded performance under 40 concurrent users, with average response times exceeding optimal thresholds.

2. Response Time Distribution
• Minimum: 4.2s (Baseline performance under load)
• Median: 19.2s (Typical user experience under stress)
• Maximum: 35.8s (Peak processing time)
• 95th Percentile: 31.0s (95% of users experience)

3. System Reliability: 96.8% Success Rate

• 121 successful requests out of 125 total
• 4 failures (3.2% error rate) indicating stress-related issues
• System maintains reasonable stability despite performance degradation

4. Throughput Analysis: 1.6 requests/second
• Processing capacity shows limitations under heavy concurrent load
• Throughput below optimal levels for production environment
• Resource contention evident from CPU and memory utilization

5. Scalability Observations
• System handling 40 users but with significant performance impact
• Error rate emergence suggests approaching system limits
• Response time variability indicates load-dependent performance degradation
• Resource utilization patterns suggest optimization opportunities

Percentile Breakdown

Percentile	Response Time (ms)	Response Time (seconds)	Assessment
50th (Median)	19,200	19.2s	■ Poor
95th	31,000	31.0s	■ Poor
99th	34,500	34.5s	■ Poor
Min	4,200	4.2s	■■ Slow
Max	35,800	35.8s	■ Very Poor

Performance Issues & Findings

■ HEAVY LOAD TEST ANALYSIS - KEY FINDINGS

1. Performance Degradation Under Load
• Average response time: 18.5 seconds (target: <2 seconds)
• 95th percentile: 31.0 seconds (indicates consistent slowdown)
• Maximum response time: 35.8 seconds (shows peak stress impact)
• **Impact:** Users experience significant delays under concurrent load

2. Error Rate Emergence: 3.2%
• 4 failed requests out of 125 total
• Error rate above optimal threshold (<1%)
• System beginning to show stress-related failures
• **Impact:** Reliability concerns under heavy concurrent usage

3. Resource Utilization Concerns
• CPU usage: 52.3% average, 100.0% peak
• Memory usage: 78.2% average, 85.4% peak
• High resource consumption with moderate throughput
• **Impact:** Efficiency issues indicating potential bottlenecks

4. Throughput Limitations: 1.6 req/sec
• Processing capacity below optimal levels for heavy load
• 125 requests processed in 78 seconds
• System struggling to maintain efficiency under concurrent stress
• **Impact:** Scalability constraints evident

5. Positive Aspects
• System maintained 96.8% success rate (reasonable reliability)
• No complete system failure or timeout
• Consistent response patterns despite degradation
• **Impact:** Core functionality remains intact under stress

Load Test Summary

```
HEAVY LOAD TEST SUMMARY:
• Test Execution: July 25, 2025 at 2025-07-25 12:38:18
• Test Duration: 78 seconds (12:37:00 - 12:38:18)
• Concurrent Users: 40 (Heavy Load Scenario)
• Total Requests: 125 (121 successful, 4 failed)
• Average Response Time: 18.5 seconds
• Throughput: 1.6 requests/second
• Success Rate: 96.8%
• ASSESSMENT: PERFORMANCE OPTIMIZATION NEEDED FOR PRODUCTION SCALE
```

Performance Optimization Recommendations

High Priority Actions

1. Response Time Optimization (Critical)
• Current average: 18.5s - Target: <2s
• Implement aggressive caching strategies for AI/ML operations
• Optimize database queries and connection management
• Consider asynchronous processing for heavy computational tasks

2. Error Rate Mitigation (High Priority)
• Current error rate: 3.2% - Target: <1%
• Investigate root cause of failures under load
• Implement circuit breaker patterns for service protection
• Enhance error handling and retry mechanisms

3. Resource Optimization (High Priority)
• CPU utilization: 52.3% average, 100.0% peak
• Memory usage: 78.2% average
• Profile and optimize resource-intensive operations
• Implement efficient memory management strategies

Medium-term Improvements

1. Scalability Enhancement
• Current throughput: 1.6 req/sec - Target: >5 req/sec
• Implement horizontal scaling capabilities
• Add load balancing for distributed processing
• Consider microservices architecture for better scalability

2. Performance Monitoring
• Establish real-time performance monitoring
• Set up alerts for response times >10 seconds
• Monitor error rates and resource utilization trends
• Implement automated performance regression testing

3. Capacity Planning
• Conduct incremental load testing (45, 50, 60 users)
• Define safe operating limits based on test results
• Establish performance baselines for different load levels
• Plan infrastructure scaling requirements

Performance vs Industry Standards

Metric	Current Performance	Industry Standard	Gap Analysis
Avg Response Time	18.5s	<2s	Exceeds by 16.5s
95th Percentile	31.0s	<3s	Exceeds by 28.0s
Throughput	1.6 req/s	>5 req/s	Below by 3.4 req/s
Success Rate	96.8%	>95%	■ Meets
Error Rate	3.2%	<1%	Exceeds by 2.2%
CPU Usage	52.3%	<70%	Exceeds by -17.7%
Memory Usage	78.2%	<80%	■ Within limits

Test Conclusion & Next Steps

The heavy load performance test conducted on **July 25, 2025 at 2025-07-25 12:38:18** with **40 concurrent users** reveals important insights about system performance under stress:

■ POSITIVE FINDINGS:
• ■ **Reasonable Reliability:** 96.8% success rate maintains core functionality
• ■ **System Stability:** No complete failures or crashes during stress test
• ■ **Consistent Behavior:** Predictable performance patterns under load
• ■ **Memory Management:** 78.2% usage within acceptable bounds

■■ AREAS REQUIRING ATTENTION:
• ■■ **Response Times:** 18.5s average exceeds optimal thresholds
• ■■ **Error Rate:** 3.2% indicates stress-related failures emerging
• ■■ **Throughput:** 1.6 req/sec below production requirements
• ■■ **CPU Utilization:** 52.3% average with 100.0% peaks showing stress

■ SCALABILITY ASSESSMENT:
The system demonstrates it can handle 40 concurrent users but with notable performance degradation. The emergence of a 3.2% error rate suggests the system is approaching its current capacity limits and requires optimization before handling higher loads.

■ BUSINESS IMPACT:
Current performance characteristics would result in:
• User experience degradation with 18.5-second average wait times
• Some users experiencing failures (3.2% error rate)
• Limited scalability for production environments requiring higher concurrency
• Need for infrastructure optimization to handle peak loads efficiently

■ RECOMMENDED ACTION PLAN:
1. **IMMEDIATE:** Performance optimization to reduce response times and error rates
2. **SHORT-TERM:** Resource utilization optimization and error handling improvements
3. **MEDIUM-TERM:** Scalability enhancements and infrastructure planning
4. **ONGOING:** Performance monitoring and capacity planning

■ SUCCESS CRITERIA FOR OPTIMIZATION:
• Average response time: <5 seconds (from 18.5s)
• Error rate: <1% (from 3.2%)
• Throughput: >5 requests/second (from 1.6 req/sec)
• CPU utilization: <70% average (from 52.3%)

■ VERDICT: OPTIMIZATION REQUIRED FOR PRODUCTION SCALE
The system shows functional capability but requires performance optimization to handle production-level concurrent loads effectively and reliably.

Report Generated: 2025-07-25 14:05:01
Source Data: ambient_api_performance_report_40users_20250725_123818.html
Test Execution: 2025-07-25 12:38:18 (12:37:00 - 12:38:18)
Analysis by: Automated Performance Testing Framework
STATUS: PERFORMANCE OPTIMIZATION RECOMMENDED