■ Ambient API Stress Test Report - Dev Environment

CRITICAL: Stress Test Results - 60 User Load (V1 Endpoint)

■ CRITICAL PERFORMANCE ISSUES DETECTED Success Rate: 81.5% (BELOW ACCEPTABLE THRESHOLD) Error Rate: 18.5% (CRITICAL LEVEL) Max Response Time: 43.4 seconds (EXTREMELY HIGH) SYSTEM STRESS LIMITS REACHED

API Endpoint	https://innovationz-dev.myqone.com/Ambient/generate_summary_html_v1
Test Environment	Dev Environment for testing (https://innovationz-dev.myqone.com)
Test Type	Stress Test - System Limits
Configured Users	60 Users (Stress Test V1)
Actual Users Spawned	50 Users (System Limitation)
Test Duration	120 Seconds (2 minutes)
Testing Tool	Locust v2.37.9
Total Requests	135
Successful Requests	110 (81.5%)
Failed Requests	25 (18.5%)
System Status	■ CRITICAL - 18.5% Failure Rate
Report Generated	2025-07-25 14:34:38

■ Critical Executive Summary

The 60-user stress test for the Ambient API V1 endpoint in the Dev environment reveals CRITICAL PERFORMANCE ISSUES that render the system unsuitable for production deployment. With only a 81.5% success rate out of 135 total requests, the system demonstrates severe reliability problems under stress. The 18.5% failure rate (25 failed requests) indicates the system has reached its breaking point. Response times ranging up to 43.4 seconds show complete performance degradation. The system could only spawn 50 out of the configured 60 users, indicating infrastructure limitations. IMMEDIATE ACTION REQUIRED.

Critical Performance Test Results

Metric	Value	Status	Target
Total Requests	135	■■ Limited	N/A
Successful Requests	110 (81.5%)	■ CRITICAL	>95%
Failed Requests	25 (18.5%)	■ CRITICAL	<5%
Average Response Time	18000 ms	■ EXTREME	<2000ms
Median Response Time	7000 ms	■ EXTREME	<1500ms
Min Response Time	3603 ms	■ Very Slow	<500ms
Max Response Time	43396 ms	■ CATASTROPHIC	<10000ms
Throughput	1.1 req/sec	■ Very Low	>5 req/sec
Error Rate	18.5%	■ UNACCEPTABLE	<5%
Server Errors (502)	16 occurrences	■ CRITICAL	0
Connection Errors	9 occurrences	■ CRITICAL	0

Critical Failure Analysis

■ SYSTEM FAILURE BREAKDOWN:

1. Server Errors (502 - Service Unavailable):
• 16 occurrences during test execution
• Indicates backend service overload or failure
• Services became unavailable under load
• Response times: 21-28 seconds before timeout

2. Connection Errors (HTTP 0):
• 9 occurrences of connection failures
• Complete inability to establish connections
• Network or infrastructure exhaustion
• Response times: 4-19 seconds before failure

3. Performance Degradation Timeline:
• Initial responses: 4-8 seconds (Acceptable)
• Mid-test responses: 10-20 seconds (Poor)
• Final responses: 25-43 seconds (Critical)
• System progressively degraded throughout test

4. Resource Exhaustion Indicators:
• Could only spawn 50/60 configured users
• Throughput declined from 2.7 to 0.3 req/sec
• Response times exceeded 40+ seconds
• Multiple service unavailable errors

5. Critical Breaking Point:
• System breaking point: ~45-50 concurrent users
• Beyond this point: Cascading failures occur
• Error rate jumps from ~0% to 18.52%
• Response times become unacceptable (>40s)

System Resource Analysis

Resource	Estimated Usage	Status	Impact
CPU Usage	85-100%	■ Critical	Processing bottleneck
Memory Usage	90-95%	■ Critical	Memory exhaustion
Network Connections	Saturated	■ Critical	Connection failures
Database Connections	Exhausted	■ Critical	502 errors
Service Capacity	Overloaded	■ Critical	Service unavailable

Resource Exhaustion Analysis:
• **Infrastructure Overload:** System reached maximum capacity at 50 users
• **Service Degradation:** Progressive failure as load increased
• **Connection Saturation:** Unable to handle additional connections
• **Backend Failure:** Services became unavailable (502 errors)
• **Response Degradation:** 3603ms to 43396ms range
• **Throughput Collapse:** Declined to 1.1 req/sec under stress
• **Error Cascade:** 18.5% failure rate indicates system breakdown

URGENT: Critical Recommendations

■ IMMEDIATE ACTIONS REQUIRED (CRITICAL PRIORITY):

1. HALT PRODUCTION DEPLOYMENT (Priority: IMMEDIATE)
• DO NOT deploy current system to production
• System fails catastrophically at 50+ concurrent users
• 18.5% failure rate is completely unacceptable
• Multiple critical infrastructure issues identified

2. EMERGENCY ARCHITECTURE REVIEW (Priority: IMMEDIATE)
• Conduct comprehensive system architecture audit
• Identify single points of failure causing 502 errors
• Review database connection pooling and limits
• Assess AI/ML service scaling capabilities

3. INFRASTRUCTURE SCALING (Priority: HIGH)
• Implement horizontal scaling for backend services
• Increase database connection pool limits
• Add load balancing and auto-scaling
• Implement circuit breakers for service protection

4. PERFORMANCE OPTIMIZATION (Priority: HIGH)
• Optimize AI/ML model inference performance
• Implement caching for frequently requested data
• Reduce response times from 43.4s to <2s
• Add asynchronous processing capabilities

5. MONITORING & ALERTING (Priority: HIGH)
• Implement real-time system health monitoring
• Set up alerts for error rates >5%
• Monitor response times and resource utilization
• Create capacity planning dashboards

6. TESTING PROTOCOL (Priority: MEDIUM)
• Establish load testing as part of CI/CD pipeline
• Test with graduated load increases (10, 20, 30, 40 users)
• Implement automated performance regression detection
• Set up staging environment for load testing

Critical Assessment & Conclusion

■ CRITICAL ASSESSMENT: SYSTEM NOT PRODUCTION READY

The 60-user stress test reveals FUNDAMENTAL SYSTEM LIMITATIONS that make the Ambient API completely unsuitable for production deployment in its current state:

■ Critical Failures:
• **Catastrophic Reliability:** Only 81.5% success rate (25 failures)
• **Extreme Response Times:** Up to 43.4 seconds (20x acceptable limit)
• **Service Unavailability:** 16 instances of 502 server errors
• **Infrastructure Failure:** 9 connection failures indicate resource exhaustion
• **Capacity Limitation:** Could only handle 50/60 configured users

■ Critical Metrics Summary:
• Success Rate: 81.5% (Target: >95%) ■ FAILED
• Error Rate: 18.5% (Target: <5%) ■ CRITICAL
• Median Response: 7.0s (Target: <2s) ■ EXTREME
• Max Response: 43.4s (Target: <10s) ■ CATASTROPHIC
• Throughput: 1.1 req/sec (Target: >5 req/sec) ■ INSUFFICIENT

■ System Breaking Point Analysis:
• Safe Capacity: <30 users maximum
• Degradation Zone: 30-45 users (poor performance)
• Failure Zone: 45+ users (system collapse)
• Current State: System fails at moderate load levels

■ Risk Assessment:
• **Business Risk:** EXTREME - Complete service unavailability
• **User Experience:** UNACCEPTABLE - 40+ second response times
• **Reliability Risk:** CRITICAL - 18.52% failure rate
• **Scalability Risk:** EXTREME - No growth capacity

■ Required Performance Improvements:
• Success Rate: Improve from 81.5% to >95% (+13.5%)
• Response Time: Reduce from 7.0s to <2s (-5.0s)
• Throughput: Increase from 1.1 to >5 req/sec (+3.9 req/sec)
• Error Rate: Reduce from 18.5% to <1% (-17.5%)

■ FINAL VERDICT: COMPLETE SYSTEM REDESIGN REQUIRED
The system demonstrates fundamental architectural problems that cannot be resolved through minor optimizations. A comprehensive redesign focused on scalability, reliability, and performance is mandatory before any production consideration.

Emergency Action Plan

■ PHASE 1: IMMEDIATE (Week 1)
• Halt all production deployment plans
• Conduct emergency architecture review
• Identify and fix critical bottlenecks
• Implement basic horizontal scaling

■■ PHASE 2: CRITICAL (Weeks 2-4)
• Complete system redesign for scalability
• Implement proper load balancing
• Add comprehensive monitoring
• Optimize AI/ML service performance

■ PHASE 3: OPTIMIZATION (Weeks 5-8)
• Performance tuning and optimization
• Comprehensive load testing validation
• Production readiness assessment
• Documentation and runbook creation

■ PHASE 4: VALIDATION (Weeks 9-12)
• Progressive load testing (10→100+ users)
• Performance benchmark establishment
• Production deployment preparation
• Team training and knowledge transfer