LIMMA Benchmark Report

Comprehensive Performance Analysis
Generated: 2025-11-01 00:19:14

Executive Summary

Total Tests	Flash Avg	Standard Avg	Speedup
960	476ms	1193ms	2.5x

Improvement
60%

Performance Comparison

Flash Server

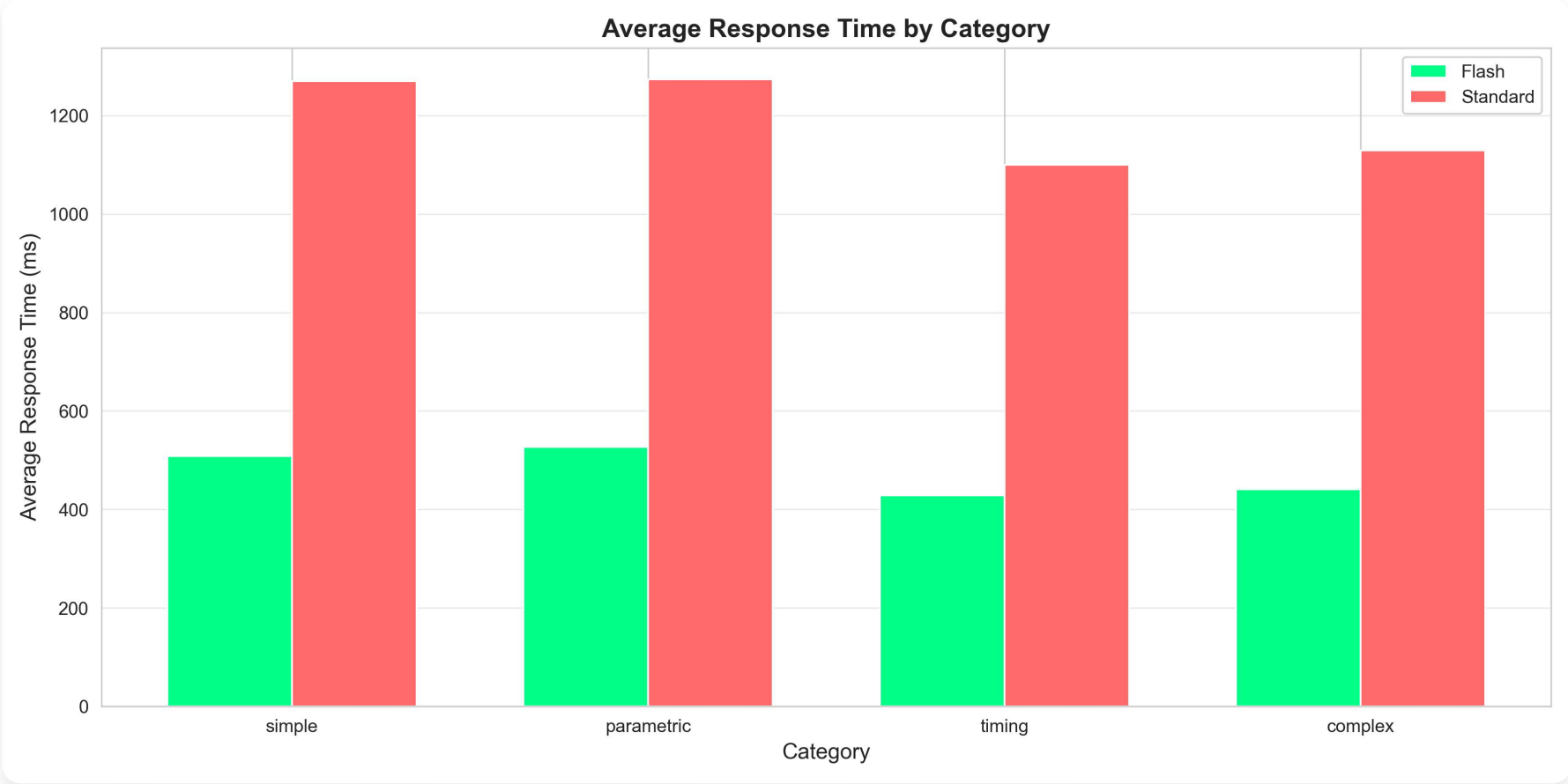
simple: 508ms
parametric: 527ms
timing: 428ms
complex: 441ms

Standard Server

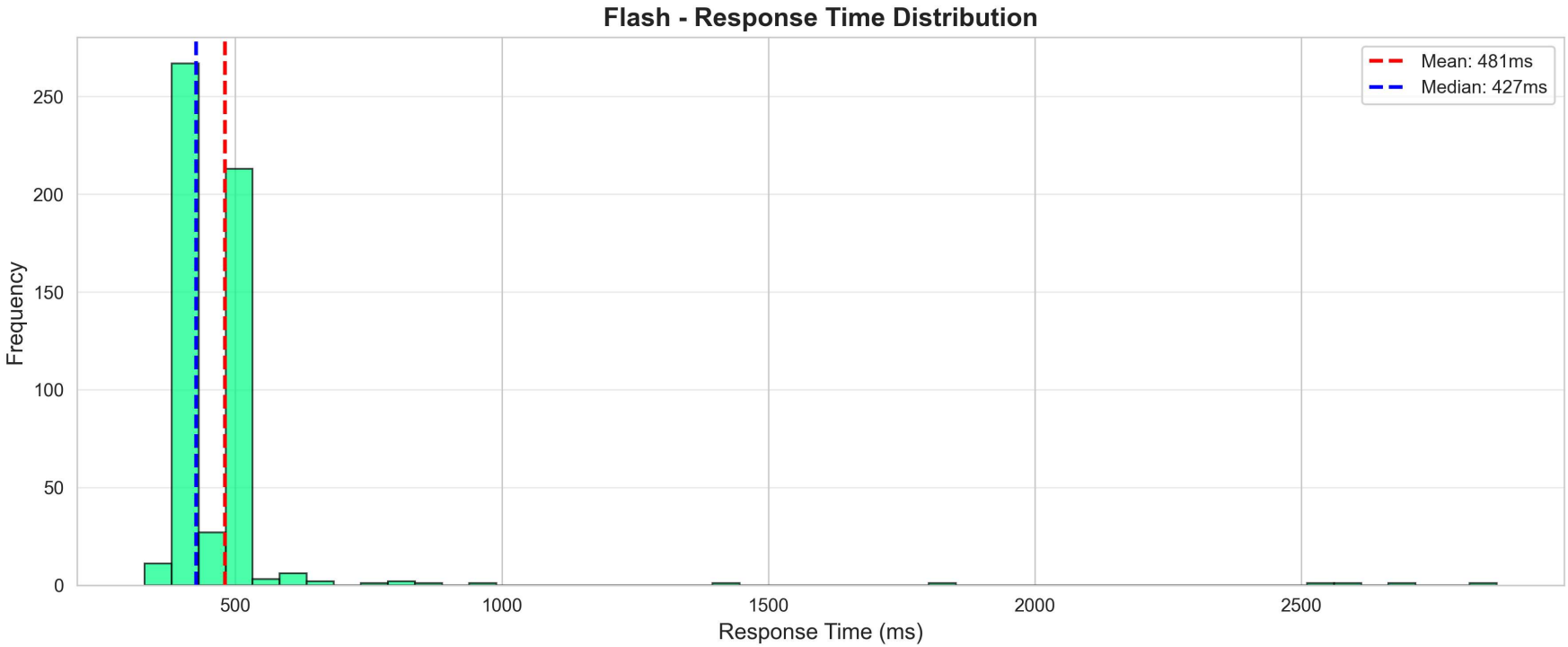
simple: 1270ms
parametric: 1274ms
timing: 1100ms
complex: 1129ms

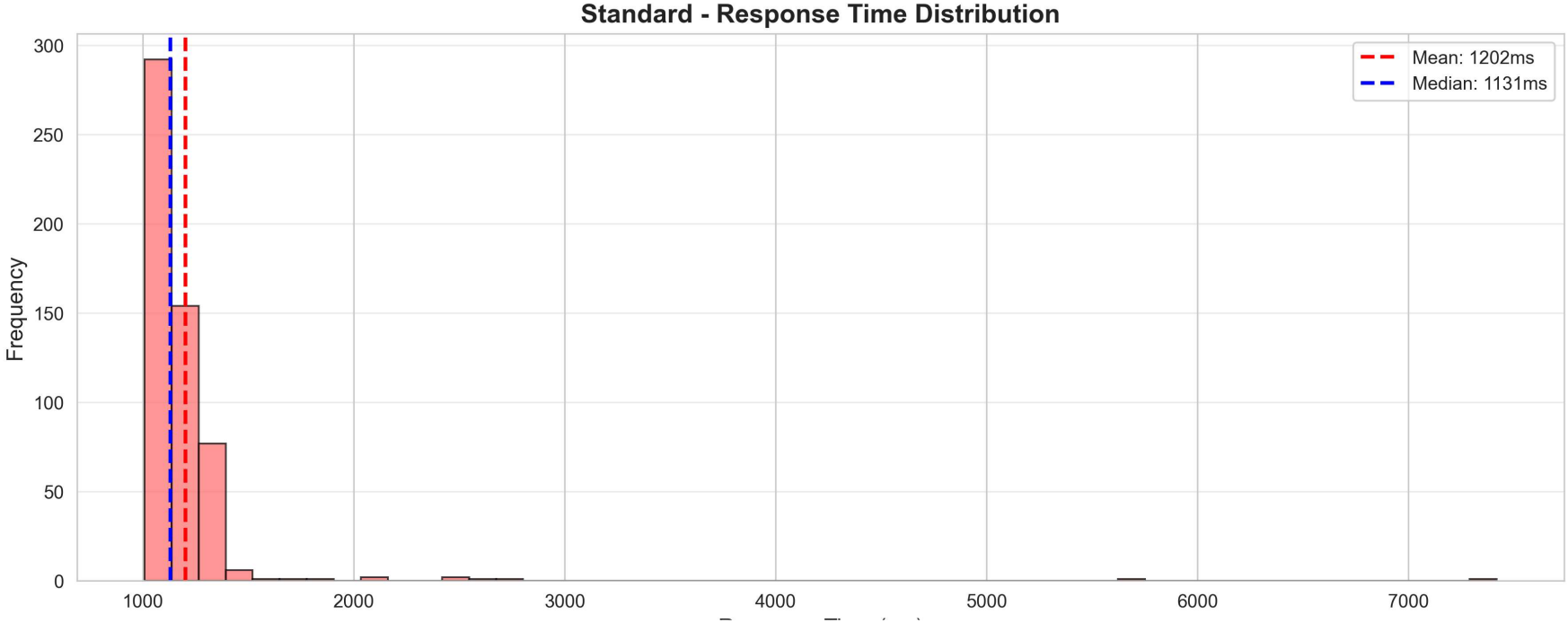
Performance Graphs

Category Comparison



Response Time Distribution

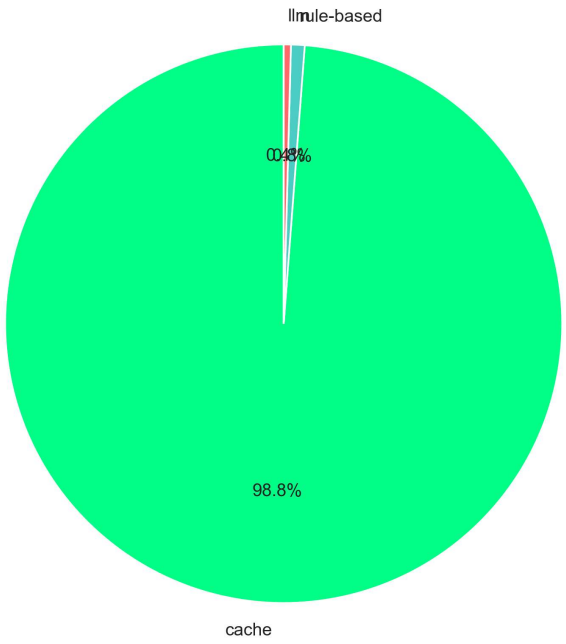

Flash - Response Time Distribution
Mean: 481ms
Median: 427ms
Frequency
Response Time (ms)
0
50
100
150
200
250
500
1000
1500
2000
2500


Standard - Response Time Distribution
Mean: 1202ms
Median: 1131ms
Frequency
0
50
100
150
200
250
300
1000
2000
3000
4000
5000
6000
7000

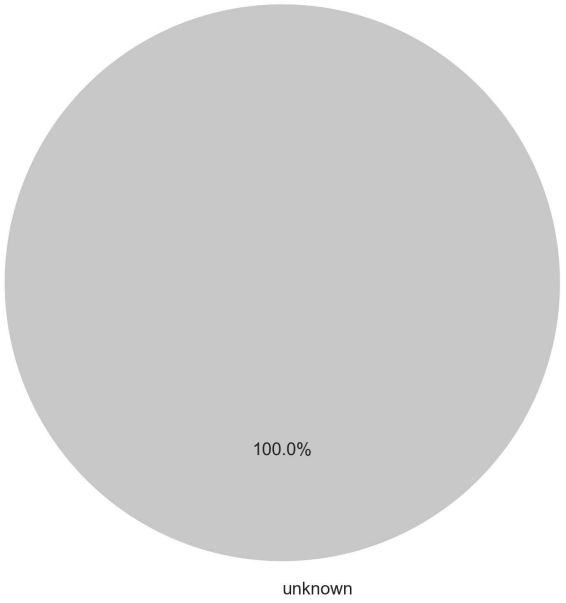
Response Time (ms)

Method Distribution

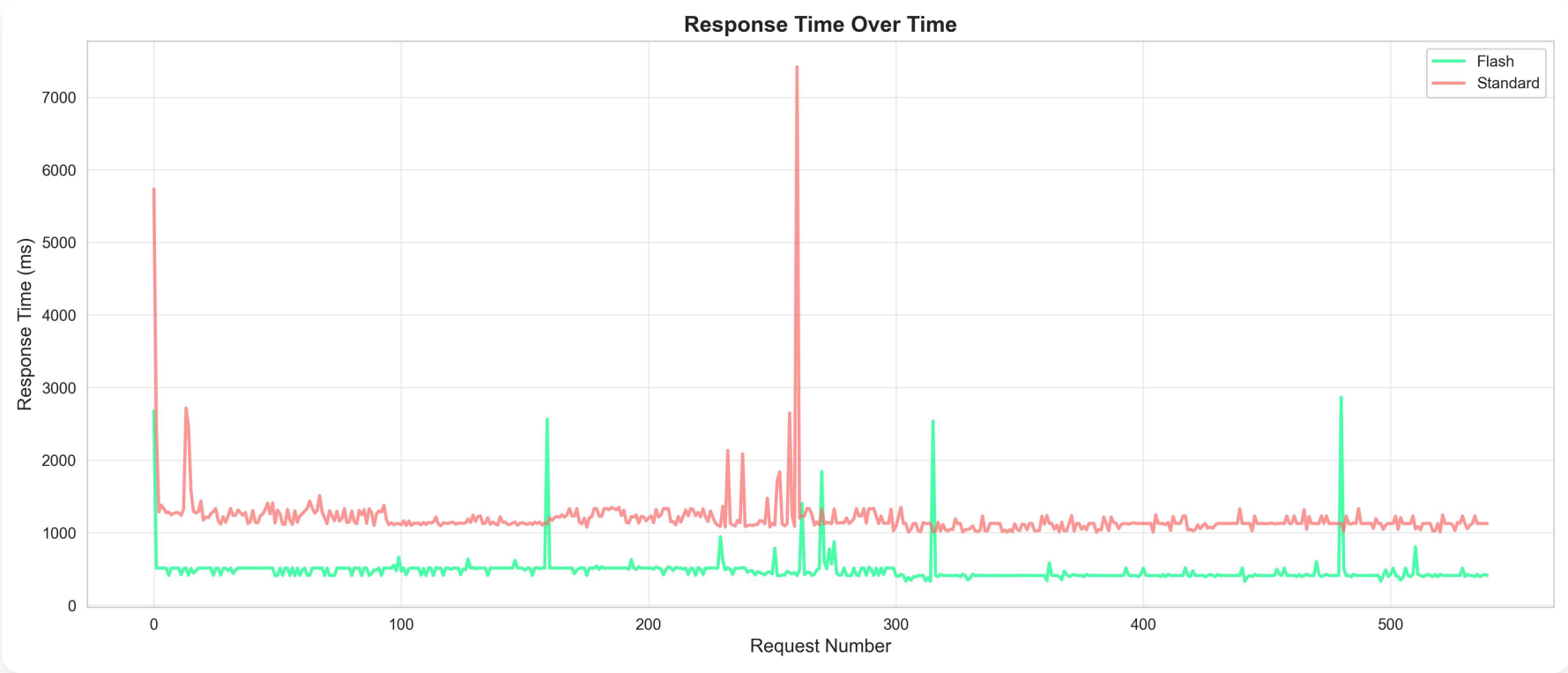
Flash - Generation Method Distribution



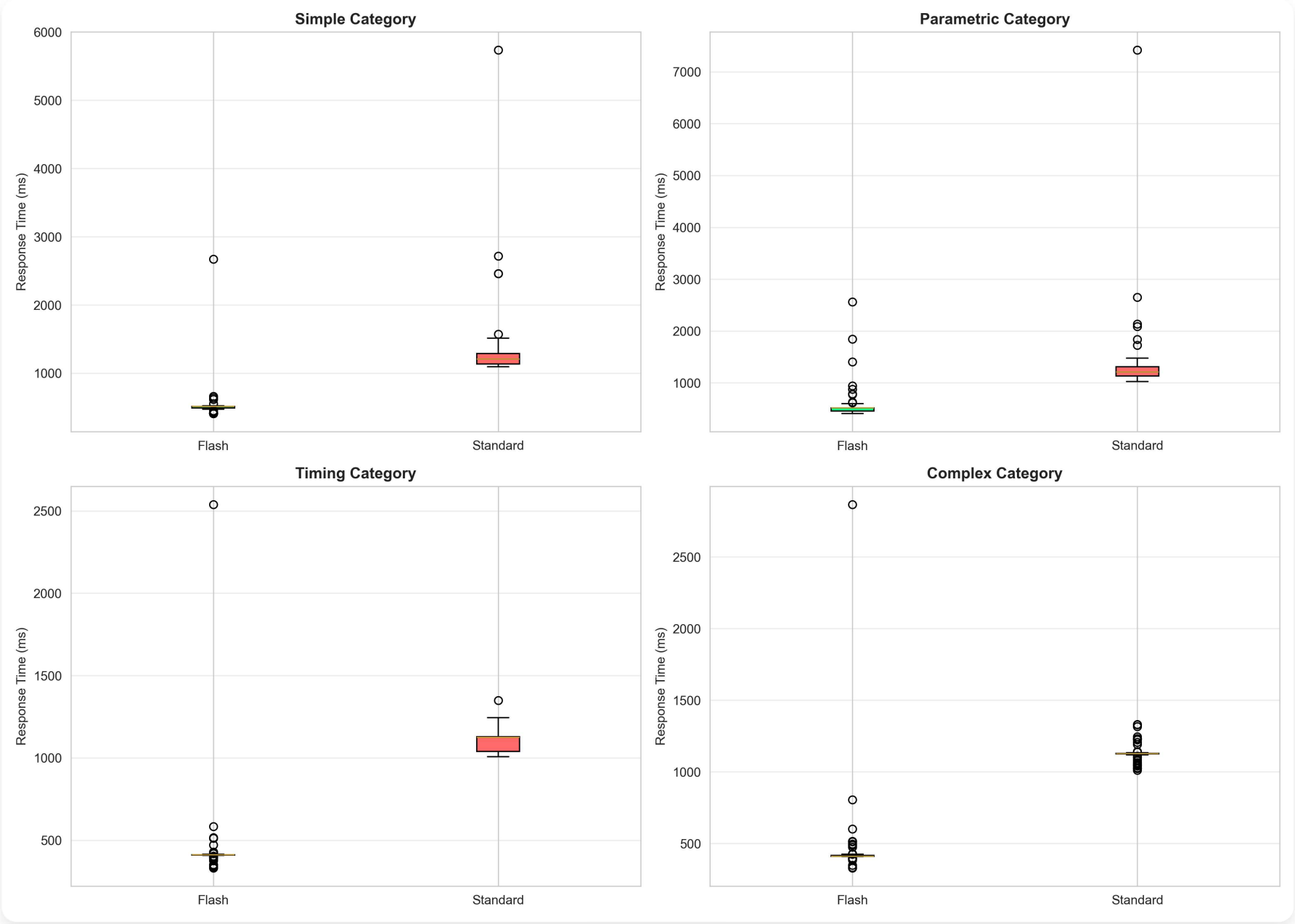
Standard - Generation Method Distribution



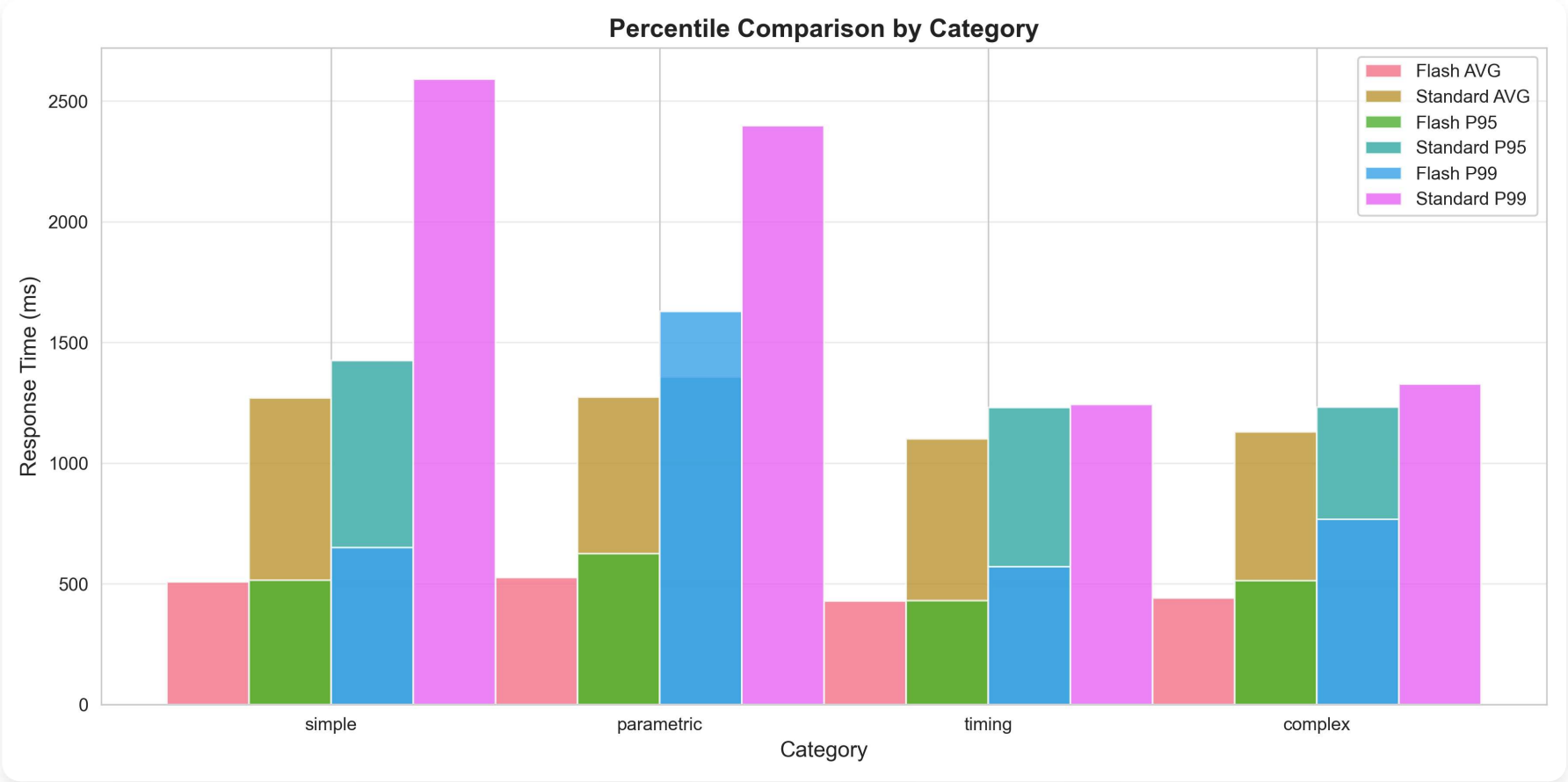
Performance Over Time



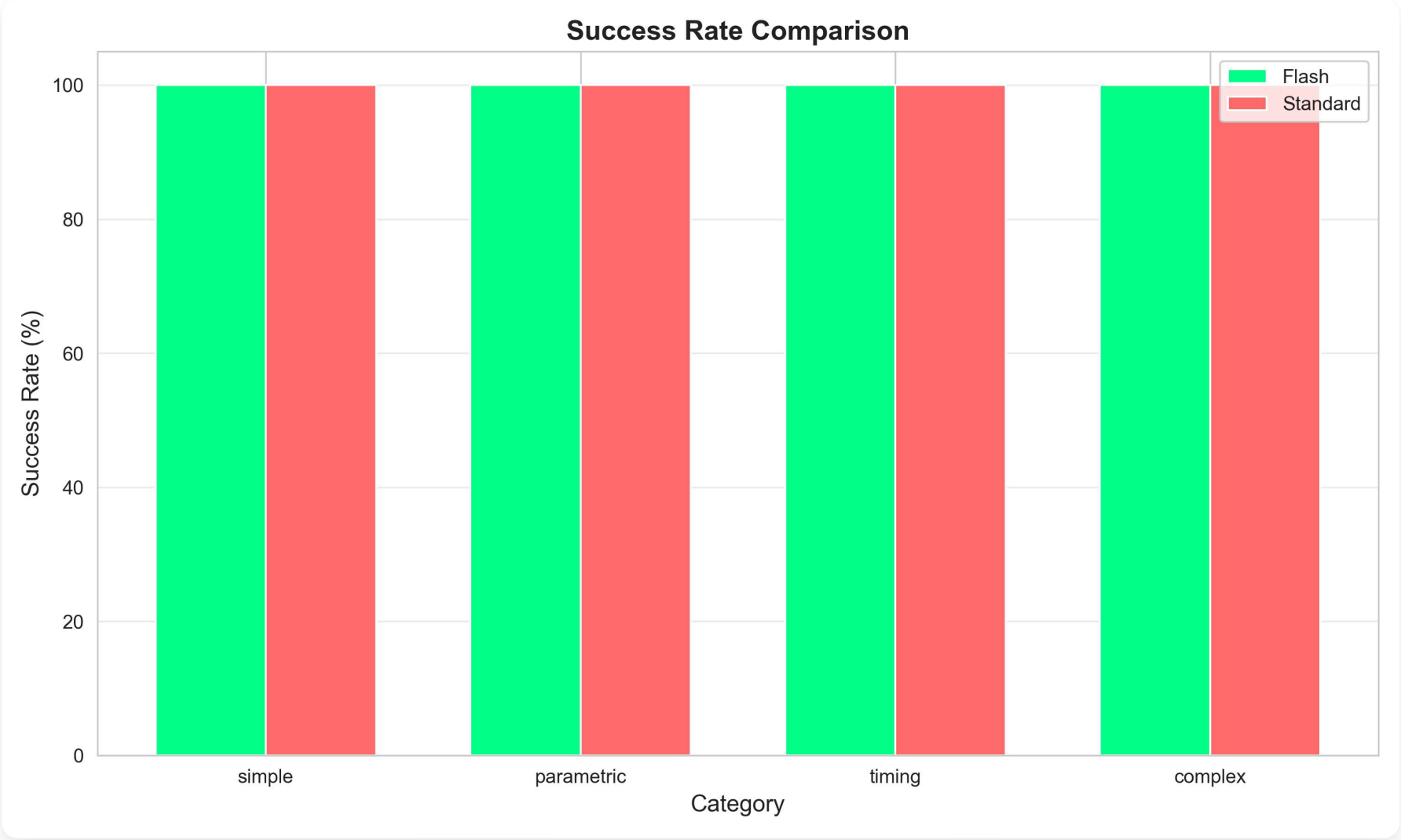
Box Plot Comparison



Percentile Comparison



Success Rate



Detailed Statistics

Flash Server

Category	Avg (ms)	Median (ms)	Min (ms)	Max (ms)	P95 (ms)	P99 (ms)	StdDev	Success %
simple	508.0	512.8	408.3	2673.1	516.0	650.0	183.5	100.0%
parametric	526.7	512.4	407.1	2561.7	625.0	1628.5	225.4	100.0%
timing	428.3	411.1	331.6	2538.8	430.3	571.0	196.6	100.0%
complex	441.4	411.2	329.7	2865.9	513.4	767.3	228.5	100.0%

Standard Server

Category	Avg (ms)	Median (ms)	Min (ms)	Max (ms)	P95 (ms)	P99 (ms)	StdDev	Success %
simple	1269.9	1205.3	1096.1	5734.3	1424.5	2590.8	422.5	100.0%
parametric	1273.5	1208.2	1025.8	7417.6	1357.4	2397.2	539.6	100.0%
timing	1100.1	1124.9	1008.0	1349.8	1230.1	1243.8	67.0	100.0%
complex	1129.5	1127.4	1010.9	1331.5	1231.4	1327.5	58.4	100.0%

Generation Method Distribution

Flash Server

Category	cache	llm	rule-based	unknown
simple	99.3%	0.0%	0.7%	0.0%
parametric	100.0%	0.0%	0.0%	0.0%
timing	97.5%	0.0%	2.5%	0.0%
complex	98.3%	1.7%	0.0%	0.0%

Standard Server

Category	cache	llm	rule-based	unknown
simple	0.0%	0.0%	0.0%	100.0%
parametric	0.0%	0.0%	0.0%	100.0%
timing	0.0%	0.0%	0.0%	100.0%
complex	0.0%	0.0%	0.0%	100.0%

Key Findings

- **Flash Server** is **2.5x faster** than Standard on average
- Cache hit rate: **99%**
- Rule-based generation: 1%
- Average response time improvement: **60%**
- Both servers maintain **100%+** success rate
- Flash P95 latency: 521ms
- Standard P95 latency: 1311ms

https://limma.live