SOFTWARE AND SCALABILITY DILIGENCE

Software Platform Scalability Review

Technical investment assessment

Investment recommendation Proceed with caution

Proceed with the current use case. The specified workload passes at 150 requests per second, compared with a near-term requirement of 80. Keep the proposed 400-request expansion outside the approved capacity assumption.

The investment question

The near-term operating plan requires 80 requests/second for a specified workload. A later expansion proposes 400. The diligence question is whether the platform supports the current investment case, and which growth assumptions should remain outside the base case.

Separate the requirement from the evidence

Capacity level	Role in the decision
80 requests/second	Near-term requirement for the specified workload
100 requests/second	Recommended initial rollout budget; an operating guardrail
150 requests/second	Demonstrated passing load in the defined replay
400 requests/second	Later expansion target; not validated

Why the recommendation is positive

Selected functional workflows and targeted authentication and authorization checks pass. Together with the passing load step, these findings support the bounded current operating plan. They do not require the investor to assume that the expansion target is already achievable.

Decision scope: the specified workload, dataset, deployment configuration, and sampled controls. The 80-request requirement is not a separately reported test result.

02 / METHOD AND PERFORMANCE EVIDENCE

What the load replay establishes

The test holds the workload definition constant

A fixed, production-shaped request mix is replayed against staging hardware and database configuration matched to the intended deployment. Dataset, tenant distribution, and request mix remain constant. Each load step has a **10-minute warm-up followed by a 45-minute measurement window**.

Acceptance requires **p95 response latency below 300 milliseconds** and an **error rate below 0.5%**. Errors include failed requests and timeouts across all attempted requests.

Reported load results

Load	p95 latency	Error rate	Assessment
150 requests/s	<300 ms	<0.5%	**Pass** Both thresholds met
250 requests/s	About 800 ms	1.8%	**Fail** Both thresholds exceeded
400 requests/s	Not tested	Not tested	Expansion target only

The degradation is material

At 250 requests per second, latency is approximately 2.7 times the 300-millisecond threshold and the error rate is 3.6 times the 0.5% threshold. The passing result at 150 therefore cannot be extrapolated to the expansion target.

The traces identify a hypothesis to isolate

Traces show database connection-pool waits alongside slow reporting queries. Their concurrence identifies a bottleneck hypothesis, not an isolated cause.

Controlled changes to one suspected constraint at a time would help isolate the cause. Application autoscaling alone is not assumed to resolve database contention.

Interpretation limit: these are discrete observations, not a capacity curve or a measured maximum. The results do not establish long-duration production behavior or performance for a different workload mix.

03 / OPERATING CONDITIONS AND INVESTOR CONCLUSION

Proceed within a defined envelope

Fund the current operating plan

The 80-request-per-second near-term requirement is supported by the passing 150-request result for the same defined workload and configuration. This is a technical basis to proceed with the current use case, with operating controls proportionate to the evidence.

Recommended rollout controls

Control	Recommended action
Initial traffic budget	Set 100 requests/second as the initial rollout guardrail.
Operating telemetry	Track p95 latency and failed requests/timeouts against the tested thresholds.
Response to degradation	Pause traffic increases and investigate sustained threshold breaches before adding load.
Change discipline	Revalidate the capacity assumption after material changes to workload mix, data shape, or deployment configuration.

The 100-request guardrail is one-third below the 150-request passing load. It is a recommended operating budget, not an additional test result or a guarantee that every production scenario below it will pass.

Make expansion a separate capital decision

Release expansion funding against a tuned design that sustains **400 requests per second** under representative conditions while meeting the same latency and error thresholds. The expansion gate applies to later growth capital, not to the decision to proceed with the current investment case.

Investor conclusion

Proceed with caution on the current use case. Base the decision on the bounded operating plan and keep unvalidated expansion capacity out of the base case. The 250-request failure creates a concrete scaling constraint; it does not negate the passing evidence for the defined near-term plan.

Limits of assurance: selected workflows and targeted access-control tests are not full-product acceptance or a comprehensive security audit. Untested functionality, security controls, production variability, and remediation cost or timing are outside this conclusion.