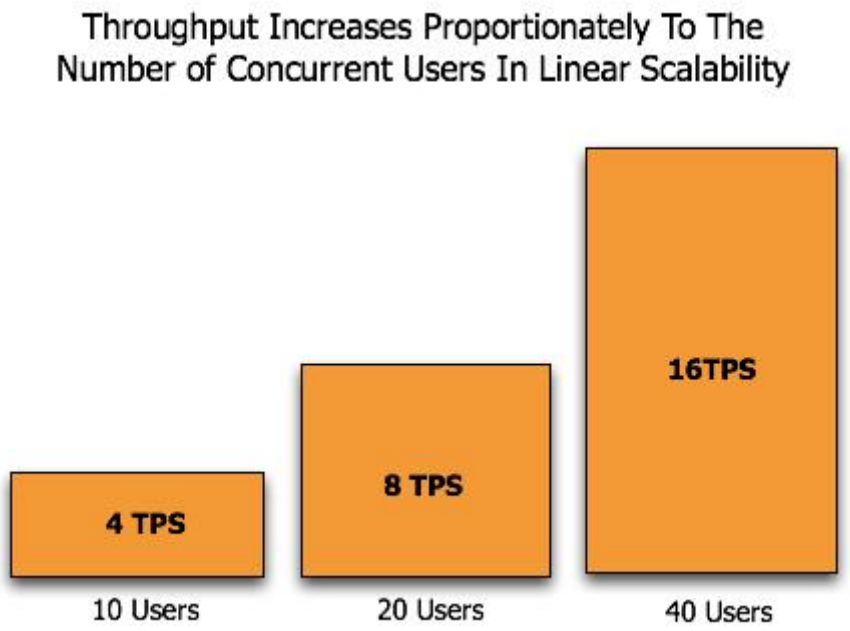
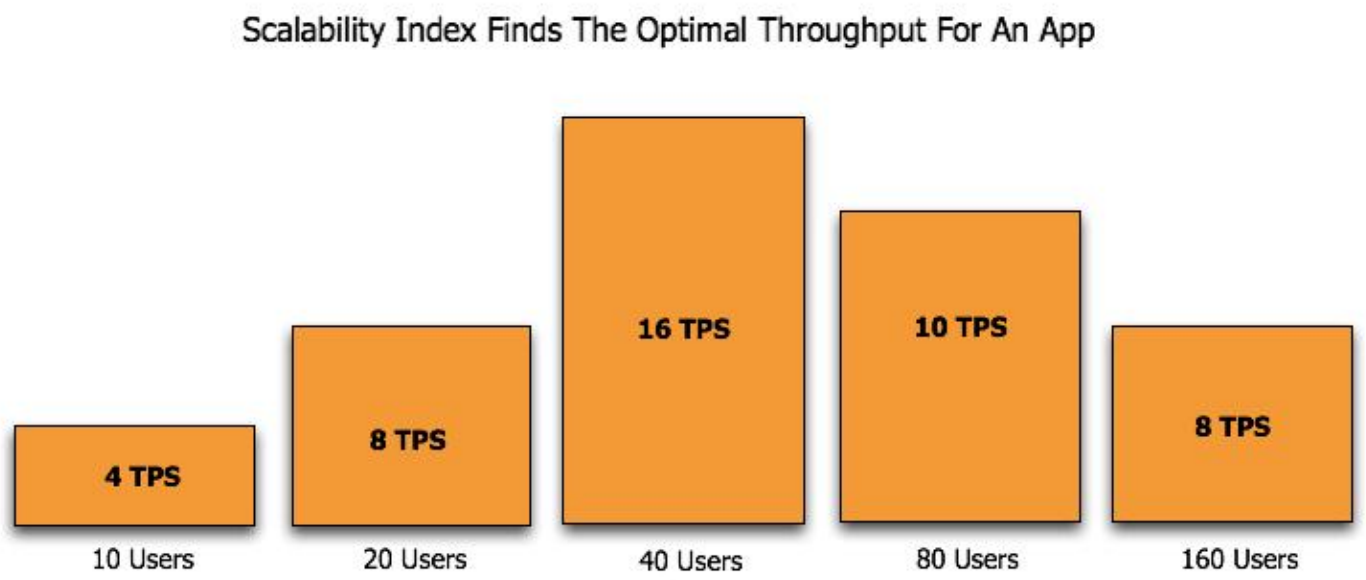
Linear Scalability in Test Servers

Start a performance test too high and the test delivers results that are a function of a performance bottleneck in the test server itself and not the application under test. Start too low and the results may not hit the upper limits of your app performance and the organization operating the test pays unnecessarily for equipment not used to its potential. Calibration testing determines the optimal level of load by identifying the topmost throughput in a test that achieves linear scalability on a single test server.



Applications with linear scalability deliver proportionally increased throughput as the test increases the number of concurrently running users. The above chart runs 10 users concurrently - with each user running a test script over and over-delivers 4 transactions per second (TPS.) A transaction is the successful operation of a test use case, test suite, or test script. With linear scalability doubling the number of users doubles the throughput. In the above example, 20 users of an application with linear scalability deliver 8 TPS. Time for a quick test: If a test at 20 virtual users delivers 8 TPS, how many TPS will 40 Users deliver with linear scalability? 40 users are double the 20 user level; therefore, 40 users deliver 16 TPS with linear scalability.

Calibration testing determines the peak throughput possible given the operating environment. A Scalability Index is a measurement of app throughput at several levels of load.



In the above example, 40 concurrent users are the optimal level for this application. When the calibration test goes to 80 users the TPS level decreases to 10 TPS. Above 80 users the application throughput continues to decrease.

There are many reasons for results as found in the above example: the app backend services may run out of network bandwidth, database caches may have filled, the load balancer may be pegging traffic to a few app servers, the app may have a logical error that prevents it from performing when more than a certain number of other users are on the app. And, of course, it could be the test server has run out of resources to operate the test!

Scaling the test to the app and the test servers is time-consuming. If each test level takes 45 minutes to operate (10 minutes to start up, 20 minutes to run, 15 minutes to ramp down) the total test time to build the above Scalability Index is 3.75 hours!

With limited performance testing time available, Appvance recommends using the Calibration Testing Methodology to rapidly identify the optimal test levels for the application under test and avoid running levels that do not lead to actionable knowledge.

Last updated: October 15, 2026

If you have any feedback on this topic, please contact us at documentation@appvance.ai

Calibration decision points

Increase load in controlled increments and compare throughput, response time, errors, and test-server resource use. Stop increasing users when the test server or application no longer scales linearly, then use the validated single-node limit when planning the distributed test.

Revision #10
Created 2026-07-31 15:09:25 UTC by Admin
Updated 2026-10-15 21:50:27 UTC by mmora