SauceLabs

Sauce Labs Continuous Testing Benchmark Report
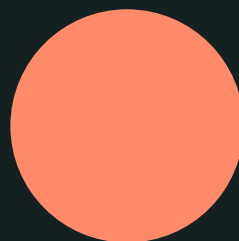
Table of Contents

SauceLabs

Executive Summary

The Sauce Labs Continuous Testing Benchmark

The Sauce Labs Continuous Testing Benchmark details the best practices critical to successful continuous testing, and tracks organizations' collective performance against those metrics based on real-world test data. The benchmark assesses performance across both desktop and mobile tests based on patterns observed in the more than five billion tests performed with the Sauce Labs unified test platform.

Benchmark Components

Test Quality
Benchmark for Excellence: Pass at least 90% of all tests run

Test Run Time
Benchmark for Excellence: Test run times averaging 2 minutes or less

Test Platform Coverage
Benchmark for Excellence: Test against at least 5 platforms on average

Test Efficiency
Benchmark for Excellence: Leverage at least 75% of available test capacity during peak testing periods

How Organizations Performed

Test Quality
% of organizations that pass at least 90% of their tests

Desktop	Mobile
24.37%	25.45%

Test Run Time
% of organizations that complete their tests in an average of two mins or less

Desktop	Mobile
45.68%	43.60%

Test Platform Coverage
% of organizations that test across five or more platforms on average

Desktop	Mobile
74.69%	63.48%

Test Efficiency
% of organizations utilize at least 75% of their available testing capacity during peak testing

Desktop/Mobile
77.96%

Just 7.79% of organizations achieved the benchmark for excellence in all four categories.

Data and Methodology

The test data used to create this report comes from the Sauce Labs Continuous Testing Cloud, the world's most comprehensive and trusted continuous testing platform. The Sauce Labs Continuous Testing Cloud is used by more than 550 enterprise organizations across the globe and spans thousands of browser/OS, mobile emulator, simulator, and real device combinations. More than three million tests are run each day on the Sauce Labs platform, and more than three billion have been run since its inception.

The report outlines the four core components of the Sauce Labs Continuous Testing Benchmark, then tracks organizations' collective performance against those metrics. The report examines the performance of mobile tests (for mobile apps and websites) in addition to desktop tests (for desktop apps and websites). The report also looks for the first time at the potential impact of failure analysis on efforts to improve test quality.

To create the report, we anonymized and analyzed Sauce Labs test data from the 30-day period. With respect to mobile tests, to ensure use-case consistency, analysis was limited to tests run using mobile emulators and simulators. Of note, though some organizations run a significantly larger number of tests on the Sauce Labs platform than others, for the purpose of this report, each organization's impact on the overall performance against a particular metric is weighted equally. This is true for both desktop and mobile test metrics.

The Path to Digital Confidence Starts With Testing

Welcome to the Sauce Labs Continuous Testing Benchmark report. We launched this report to address a significant gap in the testing market: the lack of an industry benchmark organizations could use to see how their continuous testing efforts stacked up against critical best practices and how they compared to those of other enterprises. Hundreds of organizations leveraged the report to assess the state of their continuous testing initiatives.

Since we first launched this report, the world has changed in ways no one could have imagined. By choice or necessity, every business is now a digital business, the success of which is dependent on delivering a user experience that's every bit as good as your customers expect. At Sauce Labs, we call this digital confidence. Digitally confident organizations move quickly to release new and updated applications to market, and their customers move just as quickly and confidently to use them.

Testing, of course, is the gateway to digital confidence. Companies that successfully test throughout the development cycle will grow revenue, protect their brands from the risks of a poor user experience, and deliver better products to market, faster. As is clear from the findings of this report, however, many opportunities for improvement exist, the most glaring of which is the need for improved test quality.

Though organizations did collectively improve test quality relative to 2019, we still have a long way to go. That's why the advent of failure analysis, which is discussed at length in this report, is such a promising development for the testing community, as it finally provides a means through which organizations can identify the source of common test failures and take actionable steps to improve test quality.

The discussion around failure analysis is just one way we've expanded the benchmark report. Another is the addition of mobile test data. As organizations increasingly adopt a mobile-first mindset, understanding and improving the efficacy of our mobile testing efforts is paramount. Additionally, while the parameters of the metric itself are unchanged, we have renamed what was formerly the "test concurrency" metric to "test efficiency" to better elucidate the benefits gained by fully leveraging available test capacity and running tests in parallel.

The reality is that most organizations are just beginning their continuous testing journeys, and most will face challenges along the way. We hope this report helps you identify and overcome those challenges, and serves as a vital resource on your path to digital confidence.

The Sauce Labs Continuous Testing Benchmark

Organizations focused on delivering a flawless digital experience are moving quickly to deliver automated testing throughout the software development lifecycle (SDLC). Launched for the first time last year, the Sauce Labs Continuous Testing Benchmark (CTB) defines the four best practices critical to continuous testing excellence and measures how organizations' collective testing efforts stack up against those metrics.

Though the four metrics and the associated standards of excellence that make up the CTB remain the same, the report has been extended for 2020 to examine the collective performance of organizations' mobile tests alongside that of their desktop tests. While there are certainly differences between desktop and mobile testing, the core principles of continuous testing excellence are consistent across the two categories. As such, the benchmark for excellence for each metric is the same across both desktop and mobile tests.

The Sauce Labs CTB is comprised of the following four equally weighted performance metrics, each of which is examined in greater detail later in this report:

- **Test Quality**, for which excellence is defined by passing at least 90% of all tests run
- **Test Run Time**, for which excellence is defined by running tests in an average of two minutes or less
- **Test Platform Coverage**, for which excellence is defined by testing against at least five platforms (desktop)
  or devices (mobile) on average
- **Test Efficiency**, for which excellence is defined by leveraging at least 75% percent of available test capacity during peak testing periods to drive maximum efficiency and parallelization

These metrics are averaged together to produce an overall CTB score, which ranges from 0-100. An organization that earns a score of 100 is achieving excellence across all four of these performance metrics, and is thus fully leveraging the value of continuous testing.

Benchmark Metric #1: Test Quality

There's a reason this discussion begins with an examination of test quality. Though all four components of the benchmark are of critical importance and are thus equally weighted in determining an organization's overall CTB score, there is a strong argument to be made that test quality is the single most important indicator of continuous testing health and efficacy. The other best practices that comprise the benchmark—keeping tests short, testing across a wide range of browser/OS/device combinations, and leveraging the power of parallel testing—are most effective when built on a foundation of strong test quality. Everything starts there.
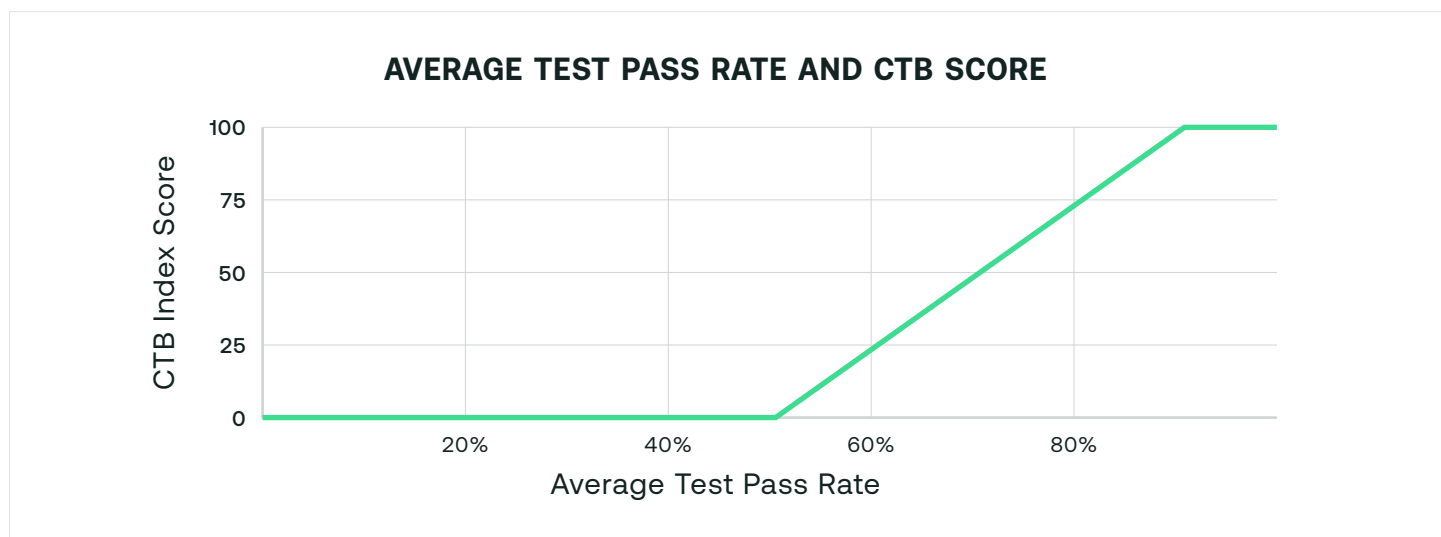
That's why it was particularly instructive to see just how much organizations struggled with test quality when the Sauce Labs CTB debuted last year. And while the numbers are improved in this year's report (as detailed below), they're still far from optimal, and highlight the glaring need for solutions that can help drive improvement.
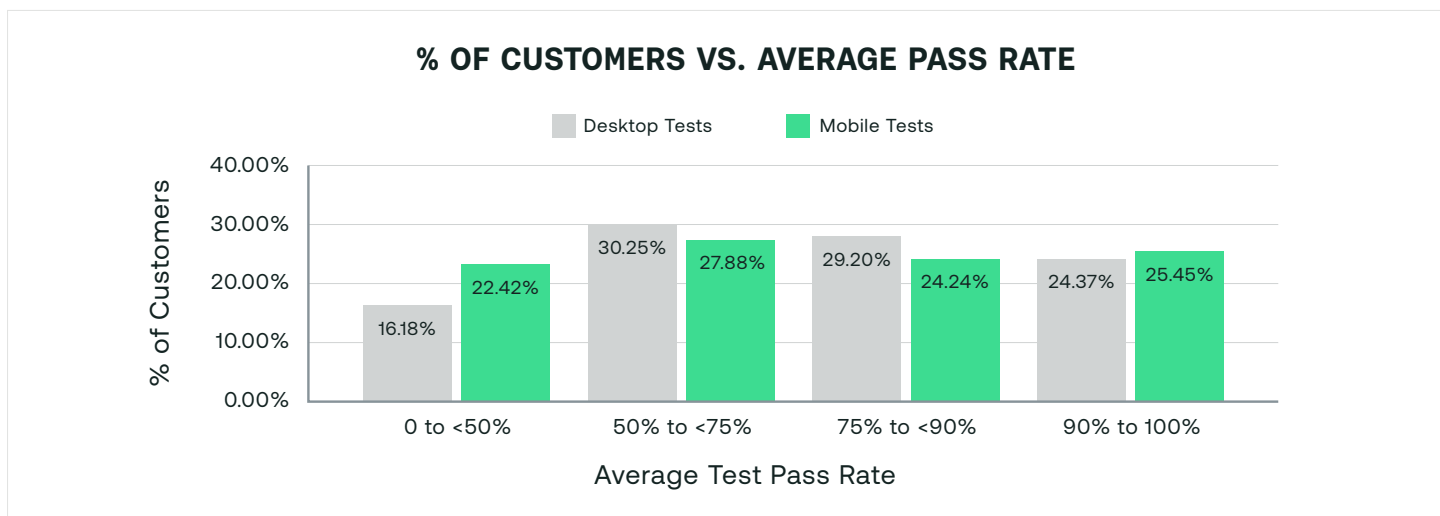
Why it matters: Remember, the purpose of automated testing is to accelerate application development and shorten the release cycle. For that to happen, the overwhelming majority of tests need to pass. When a test does fail, developers need the ability to quickly identify and fix the source of the failure. The alternative scenario, in which "flakiness" in a test suite causes a high number of failures and forces teams to manually follow-up to determine the source of those failures, is simply not conducive to delivering quality applications at speed. Testers should strive to have as low a failure rate as possible, such that failed tests carry a high signal-to-noise ratio, and developers can feel confident that a failed test has truly detected a breaking change as opposed to the "flaky" noise of an unreliable test suite.

The Benchmark for Excellence: The CTB Test Quality metric looks at the typical pass rate of an organization's tests. It's worth reiterating that the primary pain point of a failed test isn't necessarily the failure itself, but the manual follow up it necessitates. More relevant than a predefined target pass rate is the extent to which an organization has the bandwidth and resources to manually follow up on failed tests, and do so quickly. Once the number of failed tests exceeds an organization's bandwidth to quickly follow up on those failures, release velocity begins to stall.

Unfortunately, it doesn't take long for most organizations to reach that breaking point. Therefore, to achieve a test quality score of 100 signifying excellence, the average pass rate of all tests should be at or above 90%. This is consistent across both desktop and mobile tests. It should be noted that for smaller organizations with greater bandwidth constraints, pass rates may need to be even higher. Pass rates at or below 50% get a score of 0, signifying that a failed test happens just as often as a passing test.

While it's worth noting that not all failed tests are a bad thing (after all, the reason you run tests in the first place is to discover bugs and fix them before an application reaches production and the user experience is compromised), double-digit failure rates signify that an organization's test suite is not keeping pace with the application changes it's pushing. This indicates that poor test quality is undermining the organization's efforts to accelerate the delivery of flawless applications.
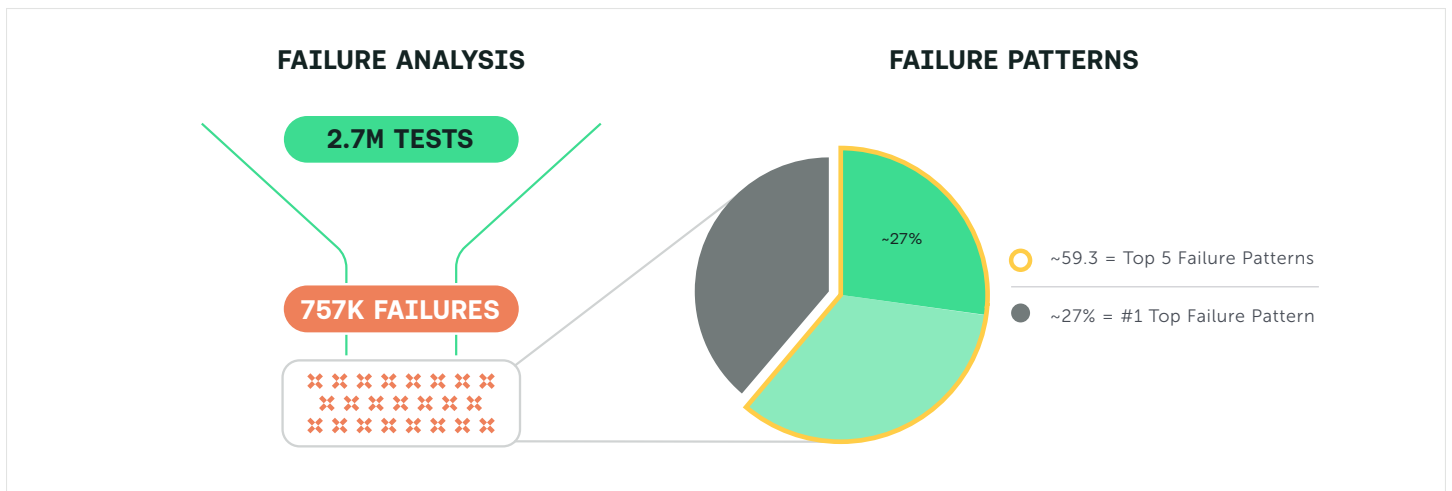
AVERAGE TEST PASS RATE AND CTB SCORE

% OF CUSTOMERS VS. AVERAGE PASS RATE

Desktop Tests | Mobile Tests

Chart: % of Customers (Y-axis, 0.00% to 40.00%) vs. Average Test Pass Rate (X-axis)

Average Test Pass Rate	Desktop Tests	Mobile Tests
0 to <50%	16.18%	22.42%
50% to <75%	30.25%	27.88%
75% to <90%	29.20%	24.24%
90% to 100%	24.37%	25.45%

How Organizations Performed: Only 24.37% of organizations pass at least 90% of their desktop tests, while 16.18 of organizations pass less than 50% of their desktop tests. Similarly, only 25.45% of organizations pass at least 90% of their mobile tests, while 22.42% of organizations pass less than 50% of their mobile tests.

The Takeaway: Though the desktop numbers represent improvement over 2019 (when they were 18.75% and 16.67%, respectively), the data for this report demonstrate a clear and increasingly urgent need for organizations to improve test quality by better understanding and remedying the source of failures. This type of proactive test-suite maintenance has long been a considerable challenge for even the most mature development teams. With the advent of failure analysis, however, organizations now have a new weapon in the battle against poor test quality.

Addressing Test Quality With Failure Analysis

Poor test quality ultimately undermines efforts to deliver quality applications at speed and grow digital confidence. Successive years of benchmark data have now shown that most organizations fail a higher-than-acceptable percentage of tests, highlighting a pressing need for improved test quality. Unfortunately, the task of maintaining a high-quality test suite has only become more difficult as organizations accelerate the pace of development and automate testing at scale. To date, there have been few levers available for organizations to pull in an effort to improve test quality, often frustrating developers as they balance the demand for faster deployments while also trying to maintain a reliable test suite. The need for new solutions to address the challenge of test quality is apparent.

One such solution is failure analysis, a new analytics-driven approach that applies sophisticated machine learning to pass/fail data in order to surface the most common reasons a given set of tests fails. With failure analysis, developers can for the first time get rapid insight into how often the same type of failure repeats itself across a given test suite. By finding and aggregating the most common causes of test failures in this manner, developers can quickly move to address the most pervasive issues and see real improvements in test quality.

FAILURE ANALYSIS

FAILURE PATTERNS

2.7M TESTS

757K FAILURES

~27%

○ ~59.3 = Top 5 Failure Patterns

● ~27% = #1 Top Failure Pattern

As part of the Continuous Testing Benchmark, Sauce Labs partnered with three enterprise organizations with highly scaled test environments on a pilot study to examine the potential efficacy of failure analysis. Using machine learning to examine 757,000 failed tests during a 30-day period, the pilot study revealed that the single most common failure pattern was responsible for 27% of all failures. In other words, the same one error, made repeatedly, was responsible for more than a quarter of all test failures. Extending the view further, the top three most common failure patterns account for 48.2% of all failures, while the top five most common failure patterns account for 59.3% of all failures.

These findings are significant for two reasons. First, they confirm that a significant percentage of test failures can be traced back to a small number of underlying issues that permeate the test suite. More importantly, they confirm that machine learning can be used to quickly surface those issues to testers and developers, who can in turn take quick action to remedy a considerable portion of their failed tests. One fix can suddenly drive 25% improvement in test quality. A handful of fixes can transform an organization struggling with test quality into one meeting and exceeding the benchmark for excellence.

Though further piloting is needed to confirm the extensibility of these findings to the general enterprise population, the initial results show failure analysis holds great promise as a new tool in the testing community's collective efforts to improve test quality, increase developer productivity, and build digital confidence.

(Disclosure: Sauce Labs now offers failure analysis as part of its continuous testing cloud.)
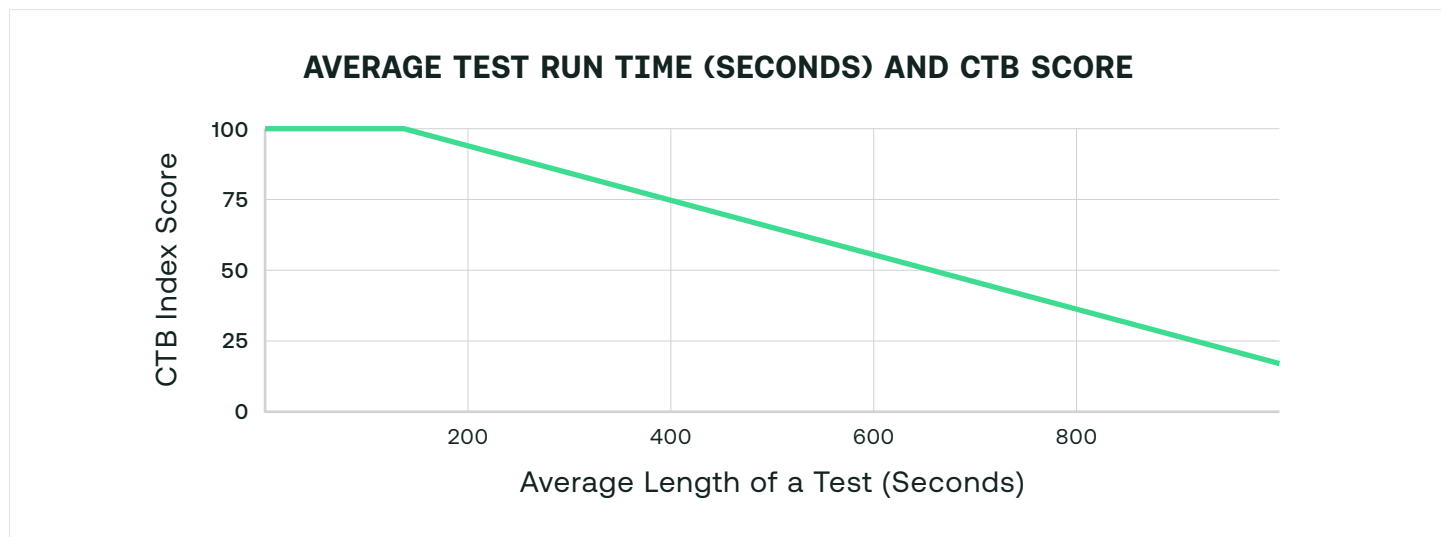
Benchmark Metric #2: Test Run Time

If test quality was the area most in need of improvement coming out of the 2019 benchmark report, test run time was right behind. In the initial report, just 35.94% of organizations met the benchmark for excellence by completing their desktop tests in an average of two minutes or less. This is significant given that one of the main reasons organizations invest in automated testing is to drive a more rapid release cycle. The longer tests take to run, the more time developers spend waiting around for feedback, and the longer it takes to push code through the delivery cycle and deliver new software to users. That's why Sauce Labs continues to encourage and advocate for the use of "atomic" tests, in which test scripts are written to evaluate a single, specific piece of application functionality, and nothing more.
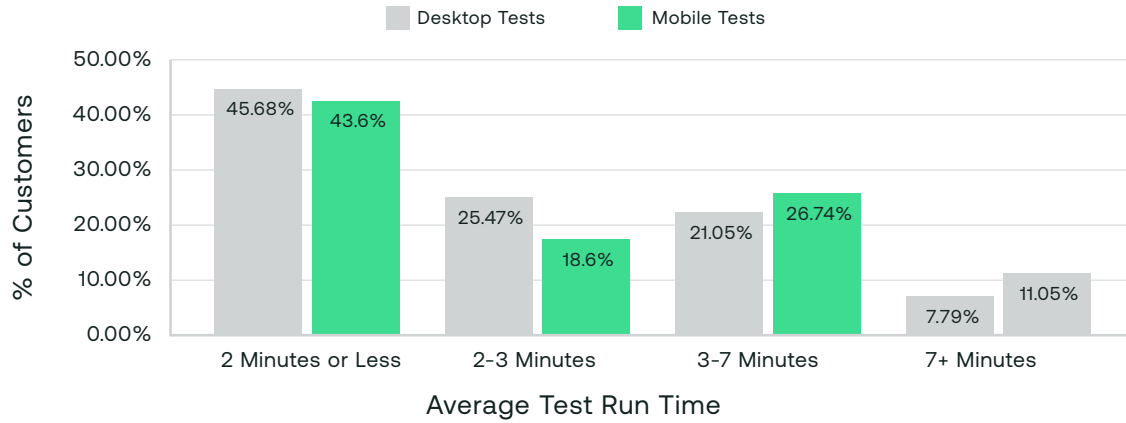
Why it Matters: For starters, as outlined above, long, slow tests do little to facilitate agile development. The benefits of keeping test run times short extend beyond just speed alone, however. As was the case last year, benchmark data for 2020 again confirms that longer tests directly lead to poor test quality, as tests that complete in two minutes or less are nearly twice as likely to pass as tests lasting longer than two minutes. Simply put, the longer a test takes to run, the more likely it is to fail.

In addition, the longer the test, the harder it is to troubleshoot in the event of failure. That's because longer tests are generally those that attempt to assess multiple pieces of application functionality with a single script. When a test of that nature fails, it becomes exponentially harder to understand precisely what's gone wrong. Counterintuitive as it may seem, it is far better to have a test suite with 100 atomic tests than a suite with 10 long tests that meander their way through multiple pieces of application functionality.

The Benchmark for Excellence: The CTB Test Run Time metric looks at the average run time of the tests organizations are running on the Sauce Labs cloud. For the reasons outlined above, to achieve a score of 100 signifying excellence, the average run time of an organization's tests should be two minutes or less.

AVERAGE TEST RUN TIME (SECONDS) AND CTB SCORE

% OF CUSTOMERS VS. AVERAGE RUN TIME

Desktop Tests ▢ Mobile Tests ▢

% of Customers

	2 Minutes or Less	2-3 Minutes	3-7 Minutes	7+ Minutes
Desktop Tests	45.68%	25.47%	21.05%	7.79%
Mobile Tests	43.6%	18.6%	26.74%	11.05%

Average Test Run Time

How Organizations Performed: 45.68% of organizations complete their desktop tests in an average of two minutes or less, while 43.60% of organizations complete their mobile tests in an average of two minutes or less.

The Takeaway: While more than half of organizations still fall short of the desired benchmark for excellence across both desktop and mobile tests, the nearly 10-point increase in the percentage of organizations completing desktop tests in an average of two minutes or less is a highly encouraging development. In the year since the launch of the first Sauce Labs CTB report, an increasing number of industry experts have publicly extolled the virtues of atomic testing, and it would appear those educational efforts are starting to pay dividends.

Benchmark Metric #3: Test Platform Coverage

As the universe of browsers, operating systems, and real mobile devices available to consumers continues to expand on a seemingly daily basis, the ability to quickly determine if an application looks, functions, and performs exactly as intended across as many of those platforms as possible is a vital component of digital confidence.

Why it matters: Today more than ever, digital platforms are the primary means—and, in many cases now, the only means—through which customers can engage with a brand and consume information, products, and services. Every business is now a digital business, and web and mobile applications are the new storefront. The more browsers, operating systems, and devices on which a business can "open" their store, the more opportunities they have to connect with customers.

The Benchmark for Excellence: The CTB Test Platform Coverage metric looks at the number of platforms against which an organization tests, and gives full marks if at least five platforms are included in their typical testing activity. For desktop tests, a platform is defined as any combination of an operating system and a browser. For mobile tests, a platform is defined as any device type. These definitions do not, however, distinguish between different versions of the same browser and or device type.

NUMBER OF PLATFORMS TESTED AND CTB SCORE

% OF CUSTOMERS VS. NUMBER OF PLATFORMS
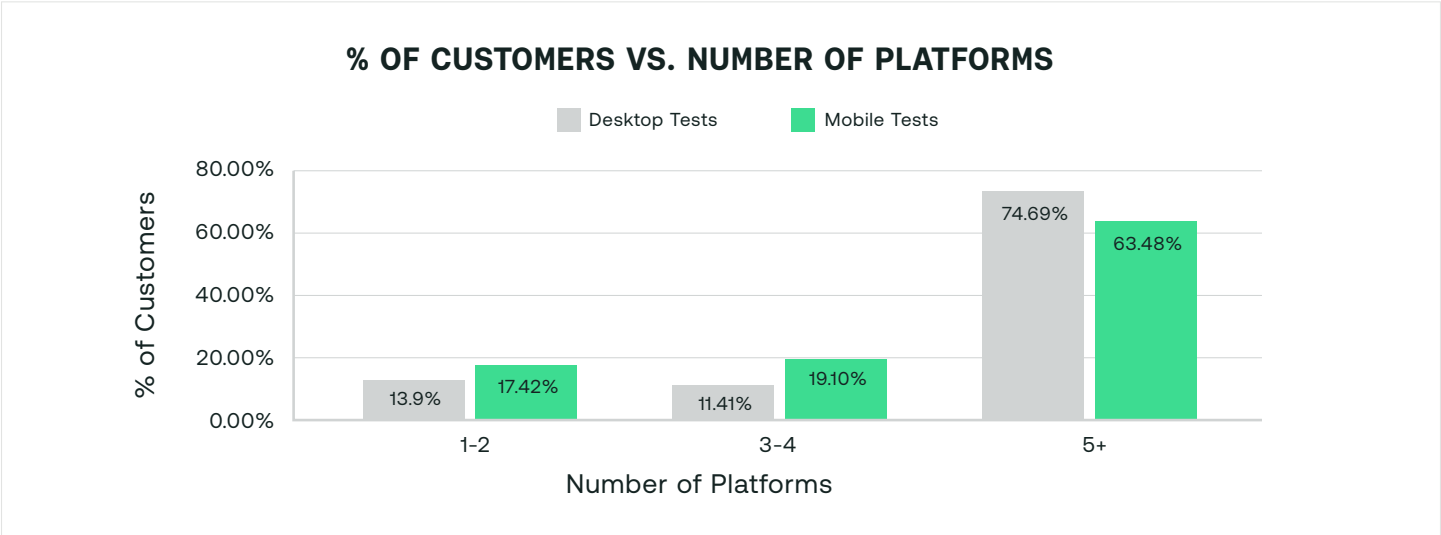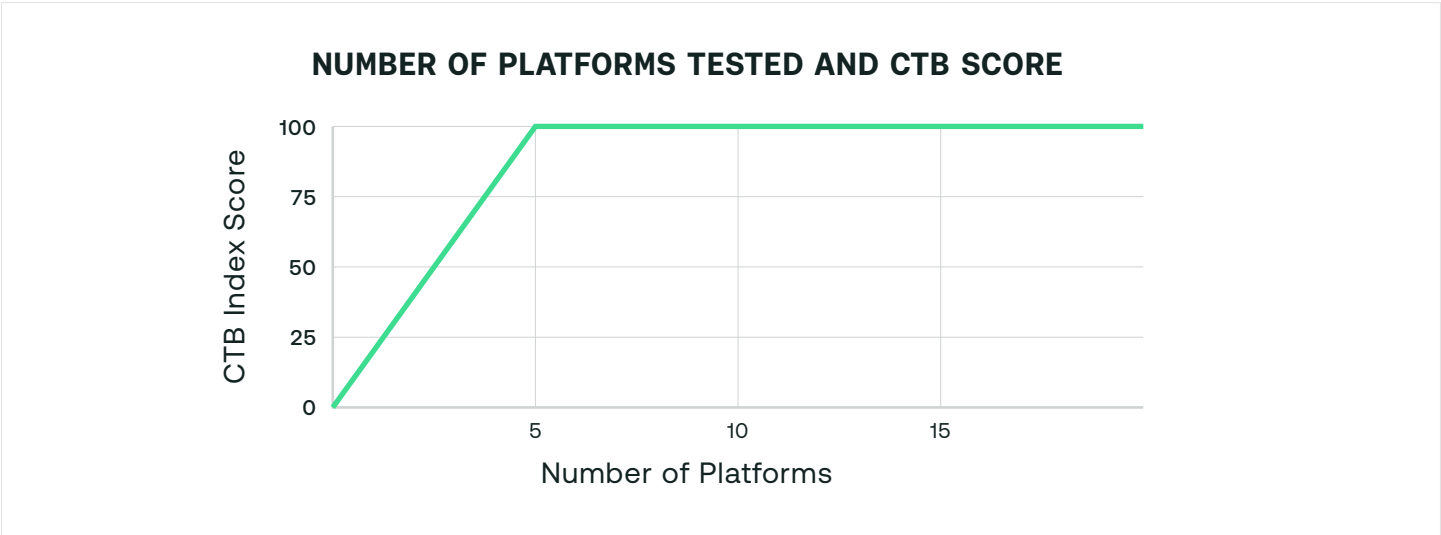
How Organizations Performed: Nearly three-quarters (74.69%) of organizations meet the benchmark for excellence with respect to desktop testing. With respect to mobile testing, the number drops slightly but not dramatically, with 63.48% of organizations testing across at least five devices on average.

The Takeaway: Enterprises have clearly gotten the memo: surviving in the all-digital era means delivering apps that work as intended whenever, wherever, and however customers wish to access them. The percentage of organizations running desktop tests across five or more platforms jumped more than 12 points (from 62.53%) in the past year, and just 13.9% of organizations still fail to test across at least three desktop platforms.

As outlined above, the numbers are slightly lower across the board with respect to mobile testing. This is not unexpected, however, given that just a small handful of makers (namely Apple and Samsung) account for the lion's share of the mobile device market. In either case, the benchmark data once again confirms that the vast majority of organizations now fully understand the importance of testing against multiple platforms.
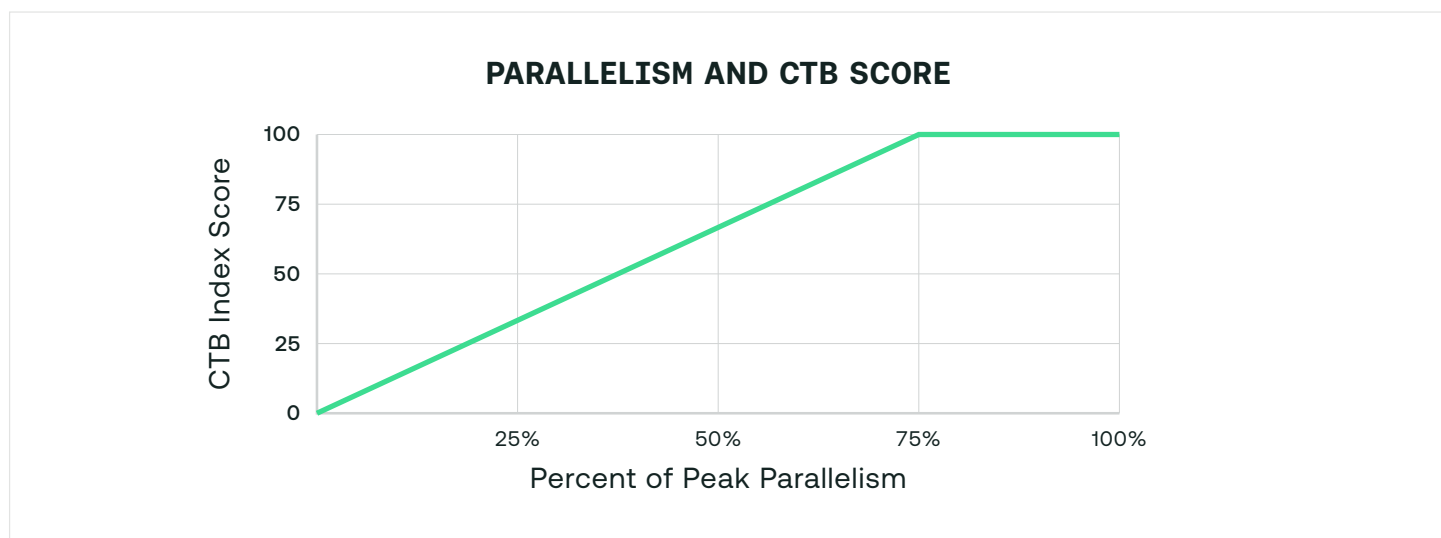
Benchmark Metric #4: Test Efficiency

Even organizations that solve for test quality, run time, and platform coverage won't enjoy the real benefits of continuous testing if they can't also solve for scale. And with enterprise development teams now often tasked with running thousands of tests each and every day, the only way to achieve that scale is through parallel testing. It is, in many ways, the final gateway to continuous testing excellence, and thus, the final metric in the Sauce Labs CTB.

Why it matters: Consider the hypothetical example of a suite of 200 tests, each of which takes one minute to run. If you can run those tests in parallel, you'll complete your entire suite of tests and have valuable feedback in the hands of your developers in just one minute. If, on the other hand, you're capacity constrained with regards to concurrency and have to run your tests sequentially, that same suite of 200 tests will now take north of three hours to complete. That's time your developers spend unproductively waiting for feedback, which eventually leads to time your customers spend impatiently waiting for your latest update or release.
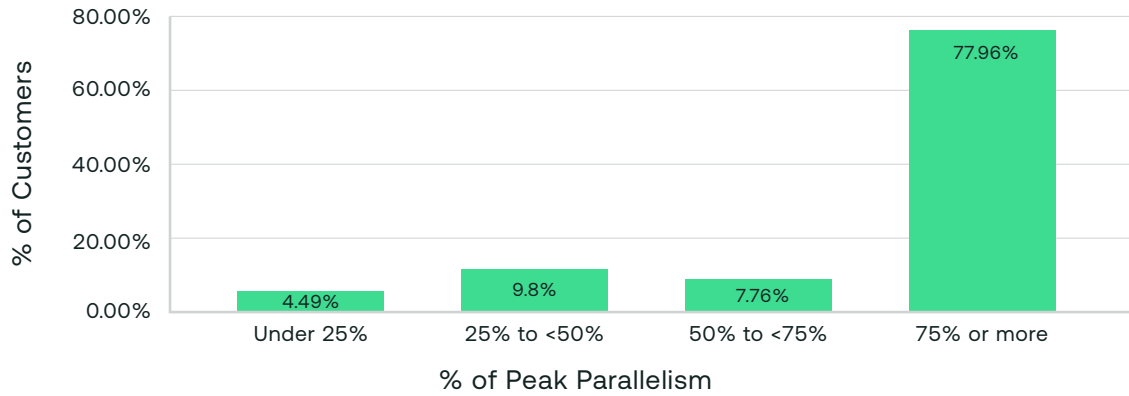
Of course, in the scenario outlined above, the organization simply does not have the additional test capacity it needs to leverage parallel testing. An equally common (and frankly more concerning) scenario is one in which an organization has the necessary test capacity but simply fails to leverage it. The testing equivalent of ordering the buffet but only eating an appetizer, failing to maximize your available test capacity is both operationally and financially inefficient.

The Benchmark for Excellence: The CTB Efficiency metric measures the percentage of provisioned testing capacity an organization utilizes when a test suite is running. The greater the percentage of test capacity leveraged, the more efficiently an organization is utilizing concurrency. To achieve a score of 100 signifying excellence, an organization should utilize at least 75% of their available test capacity during peak testing periods.

Of note, for this particular metric, we did not differentiate performance across desktop and mobile tests. For users of Sauce Labs Continuous Testing Cloud, capacity is capacity. There are no limitations, restrictions, or requirements dictating how an organization allocates their test capacity across desktop and mobile tests.

PARALLELISM AND CTB SCORE

% OF CUSTOMERS VS. % OF PEAK PARALLELISM

% of Customers (y-axis)

- Under 25%: 4.49%
- 25% to <50%: 9.8%
- 50% to <75%: 7.76%
- 75% or more: 77.96%

% of Peak Parallelism (x-axis)

How Organizations Performed: More than three quarters (77.96%) of organizations currently utilize at least 75% of their available testing capacity during peak testing periods.

The Takeaway: The percentage of organizations leveraging at least 75% of their available test capacity jumped more than seven points (from 70.88%) in the past year, again delivering solid improvement on an already strong area of performance. That said, as development processes continue to mature and organizations increasingly become more agile, there's little excuse for this number to not be at or near 100%. The roughly 15% of organizations still utilizing less than half of their available test capacity should consider this a missed opportunity, and focus on designing test environments that fully leverage their available capacity when running a test suite.
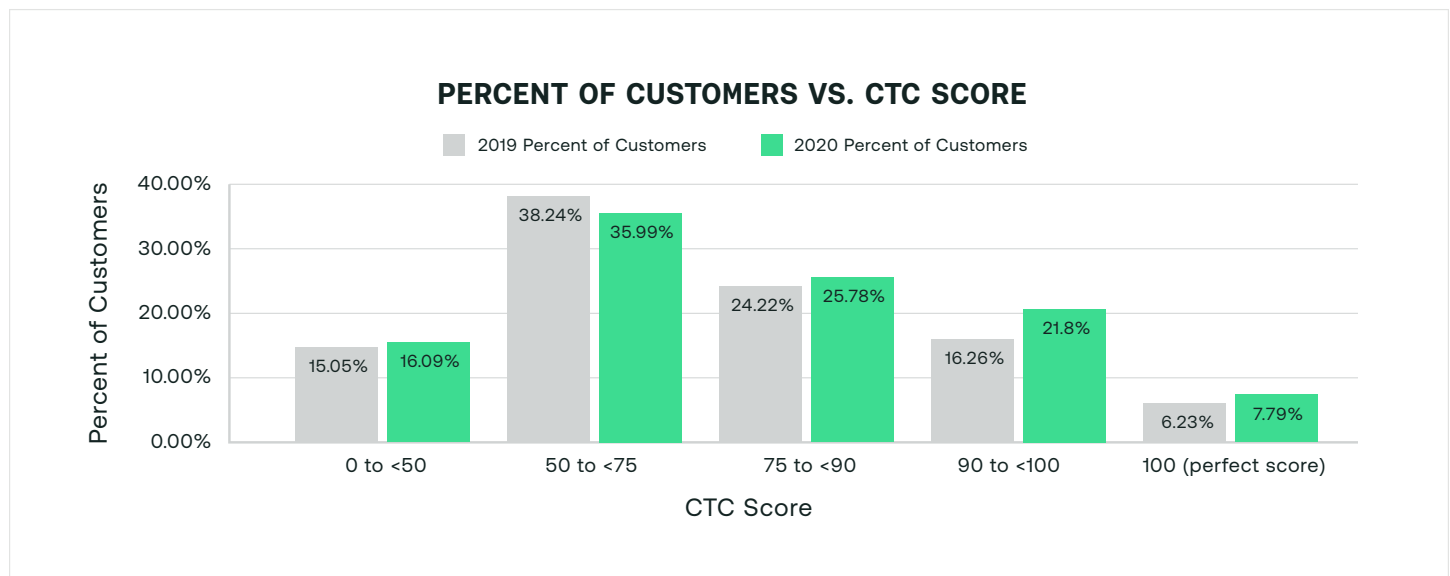
Putting It All Together: The Continuous Testing Benchmark

Combining the aforementioned metrics—test quality, test run time, test platform coverage, and test efficiency—into a single, equally weighted metric creates the overall Sauce Labs Continuous Testing Benchmark.

Why it Matters: The only resource of its kind available to the testing community, the benchmark provides organizations with a single viewpoint into the efficacy of their overall continuous testing efforts.

Benchmark for Excellence: To achieve a perfect score (100), one signifying complete and total continuous testing excellence, organizations must meet or exceed the benchmark for excellence in each of the four component categories as previously outlined.

How Organizations Performed: Only 7.79% of organizations achieved a perfect score of 100, signifying excellence across all four of the benchmark's component metrics.

PERCENT OF CUSTOMERS VS. CTC SCORE

Legend: 2019 Percent of Customers (gray), 2020 Percent of Customers (green)

CTC Score	2019 Percent of Customers	2020 Percent of Customers
0 to <50	15.05%	16.09%
50 to <75	38.24%	35.99%
75 to <90	24.22%	25.78%
90 to <100	16.26%	21.8%
100 (perfect score)	6.23%	7.79%

Y-axis: Percent of Customers. X-axis: CTC Score.

The Takeaway: The percentage of organizations achieving the benchmark for excellence across all four metrics is up from 2019 (when it was just 6.23%), but not significantly so. That less than 10% of organizations are meeting the standard of excellence in every category is again a reflection of the relatively nascent state of continuous testing as the engine that drives DevOps initiatives, and of the relative lack of testing expertise that exists in organizations large and small.

That said, there are encouraging signs. Nearly 30% of organizations (29.59%) achieved an overall benchmark score of 90 or above in this year's report, up from 22.49% in 2019, while 55.37% achieved an overall benchmark score of 75 or above, up from 46.71% in 2019. Taken together, these indicators suggest that organizations are increasingly aware of the critical role continuous testing plays in delivering quality software at speed, and are increasingly invested in learning and implementing the best practices necessary to drive improvement.

Conclusion

With two years of Sauce Labs Continuous Testing Benchmark data now available to draw on, we can see a number of encouraging trends. First and foremost, organizations continue to prioritize implementing continuous testing throughout the software development lifecycle, and continue to make considerable progress and improvements in their efforts to do so. The number of organizations achieving the standard for excellence grew by at least 5% across each of the four benchmark metrics. That mobile testing performance metrics closely mirrored desktop testing metrics is also encouraging. Despite mobile being a more nascent area of testing, organizations are doing well to carry over established best practices.

Staying on the positive side of the ledger, at least 75% of organizations are now meeting the benchmarks for excellence with respect to test efficiency and test platform coverage. As a collective group, we're testing across a wide variety of platforms and running tests in parallel as we do. In addition, although the overall numbers for test run time are not yet where they need to be, the nearly 10% improvement in the number of organizations meeting the benchmark for excellence is highly encouraging.

On the flip side, test quality remains a challenge. Though the percentage of organizations meeting the benchmark for excellence did improve (from 18.75% to 24.37%), that more than 75% of organizations are still failing more than 10% percent of their desktop tests on average is cause for concern. If left unchecked and unimproved, poor test quality can potentially undermine the progress seen in each of the other three areas.

Still, even this concerning trend is tempered by the findings of the pilot study of failure analysis, which revealed that organizations can potentially address more than a quarter (27%) of all test failures by surfacing and remedying the most common failure pattern within a given test suite. Given the urgent need to improve test quality, failure analysis stands as perhaps the most promising application of analytics and machine learning the testing industry has seen to date.

What this same set of performance metrics will look like a year from now is anyone's guess, of course, but the arrow is undoubtedly pointing in the right direction, toward a future defined by digital confidence.

Practitioner Recommendations

- Run atomic tests to improve test quality, reduce test run times, and achieve parallelization.

- Leverage failure analysis to surface your most common test errors, prioritize how frequently they are occurring, and gain insight into where you can focus efforts to improve your pass rates.

- Adopt a mobile-first strategy in which your investment in mobile testing and development is proportional to current and anticipated customer usage.

- Invest in (and use!) the testing capacity necessary to deliver automation at enterprise scale.

- Seek the help of experts. Testing is hard. Experienced testing experts can help you navigate the challenges.

How is my organization doing?
Want to know how the tests your team is running perform against the Sauce Labs Continuous Testing Benchmarks?
If you are a current Sauce Labs customer and would like a custom report showing your Continuous Testing Benchmark scores and recommendations, please contact your Customer Success Manager. If you're not a Sauce Labs customer but would like to learn more about improving your overall continuous testing performance, contact us here.
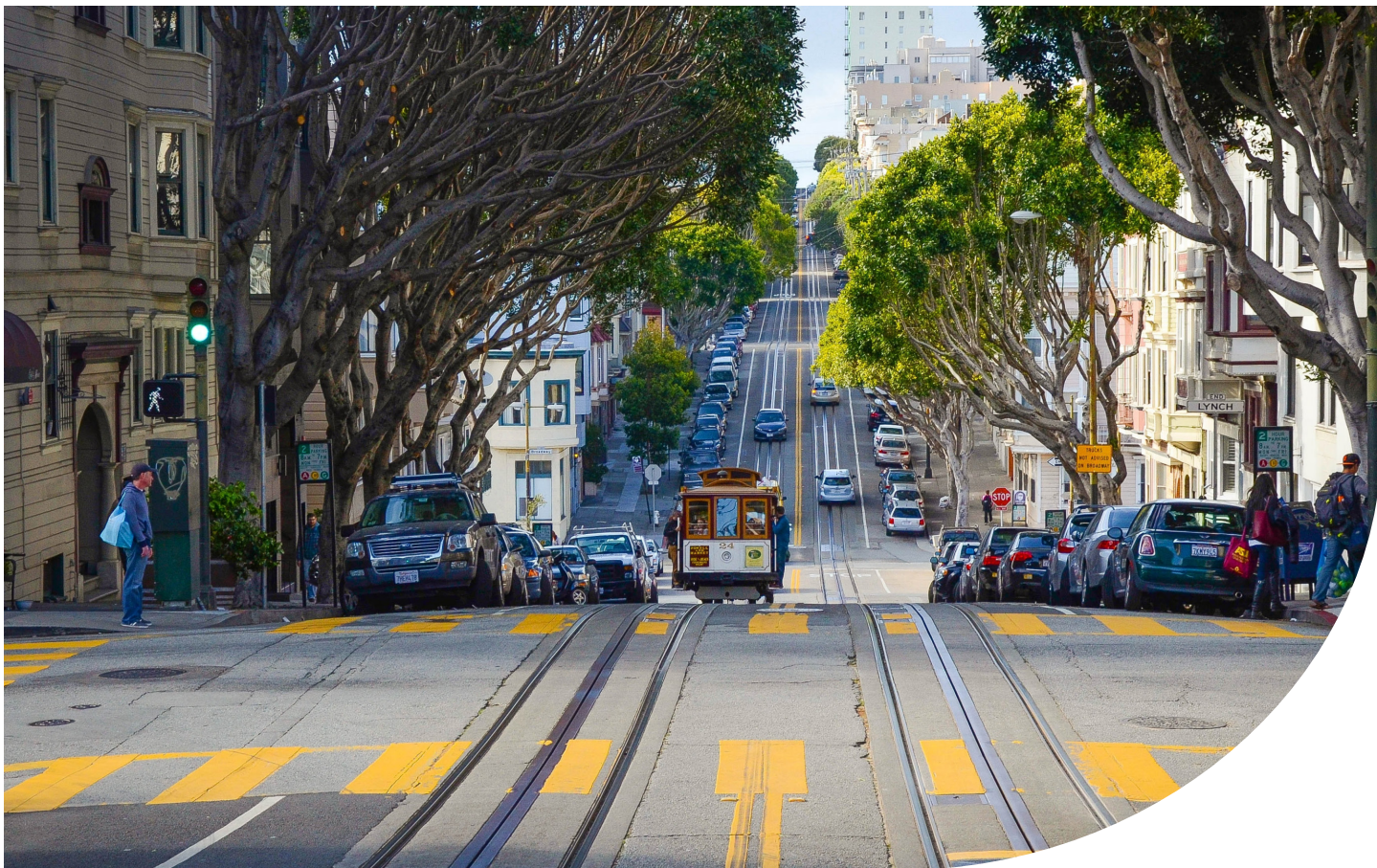
About Sauce Labs

Sauce Labs is the leading provider of continuous test and error reporting solutions that give companies the confidence to develop, deliver and update high quality software at speed. The Sauce DevOps Test Toolchain identifies quality signals in development and production, accelerating the ability to release and update web and mobile applications that look, function and perform exactly as they should on every browser, operating system and device, every single time. Sauce Labs is a privately held company funded by TPG and Riverwood Capital.

For more information, please visit

→ **saucelabs.com**

SauceLabs