(benchmarks)=

# Benchmark Suites

vLLM contains two sets of benchmarks:

- [Performance benchmarks](#performance-benchmarks)
- [Nightly benchmarks](#nightly-benchmarks)

(performance-benchmarks)=

## Performance Benchmarks

The performance benchmarks are used for development to confirm whether new changes improve performance under various workloads. They are triggered on every commit with both the `perf-benchmarks` and `ready` labels, and when a PR is merged into vLLM.

The latest performance results are hosted on the public [vLLM Performance Dashboard](https://perf.vllm.ai).

More information on the performance benchmarks and their parameters can be found [here](gh-file:.buildkite/nightly-benchmarks/performance-benchmarks-descriptions.md).

(nightly-benchmarks)=

## Nightly Benchmarks

These compare vLLM's performance against alternatives (`tgi`, `trt-llm`, and `lmdeploy`) when there are major updates of vLLM (e.g., bumping up to a new version). They are primarily intended for consumers to evaluate when to choose vLLM over other options and are triggered on every commit with both the `perf-benchmarks` and `nightly-benchmarks` labels.

The latest nightly benchmark results are shared in major release blog posts such as [vLLM v0.6.0](https://blog.vllm.ai/2024/09/05/perf-update.html).

More information on the nightly benchmarks and their parameters can be found [here](gh-file:.buildkite/nightly-benchmarks/nightly-descriptions.md).