This page describes the CPU utilization metrics that Cloud Spanner provides. You can view these metrics in the Google Cloud Console (/spanner/docs/monitoring-console) and in the Stackdriver Monitoring console (/spanner/docs/monitoring-stackdriver).

When Cloud Spanner measures CPU utilization, it organizes tasks into the following categories:

	User tasks	**System tasks**
High priority	**High-priority user tasks**	**High-priority system tasks**
	Tasks that your application initiates and that Cloud Spanner handles as a high priority. Includes most read and commit requests. Also includes parts of the work for batch writes, but not batch reads.	Tasks that Cloud Spanner initiates and handles as a high priority. Includes backfilling an index and data splitting.
Low priority	**Low-priority user tasks**	**Low-priority system tasks**
	Tasks that your application initiates, and that do not need to be completed as quickly as high-priority tasks. Includes batch reads and batch queries. Also includes parts of the work for batch writes.	Tasks that Cloud Spanner initiates, and that do not need to be completed as quickly as high-priority tasks. Includes database compaction and schema change validation.

High-priority tasks immediately preempt low-priority tasks. If necessary, Cloud Spanner stops all low-priority tasks and allows high-priority tasks to utilize up to 100% of the available CPU resources. While low-priority system tasks can be delayed in the short term, they must run eventually for optimal performance. Therefore, **you must provision your instance with enough nodes** (#reduce) **to handle both high- and low-priority tasks**.

If there are no high-priority tasks, Cloud Spanner will utilize up to 100% of the available CPU resources to complete low-priority tasks more quickly. Spikes in background usage are not a

sign of a problem. Low-priority tasks can yield to high-priority tasks, including user tasks, almost instantly.

Cloud Spanner provides the following metrics for CPU utilization:

- **Rolling average 24 hour**: A rolling average of total CPU utilization, as a percentage of the instance's CPU resources, for each database. Each data point is an average for the previous 24 hours.
- **High priority**: The CPU utilization, as a percentage of the instance's CPU resources, for high-priority tasks.
- **Total**: The total CPU utilization, as a percentage of the instance's CPU resources.

  For *instances*, you can view total CPU utilization by database or by task priority.

  For *databases*, you can view total CPU utilization by task priority.

You can view charts for these metrics in the Cloud Console (/spanner/docs/monitoring-console) or in the Stackdriver Monitoring console (/spanner/docs/monitoring-stackdriver). You can also use the Stackdriver Monitoring console to create alerts for high CPU utilization (/spanner/docs/monitoring-stackdriver#create-alert), as described below.

The following table specifies our recommendations for maximum CPU usage for both single-region and multi-region instances. These numbers are to ensure that your instance has enough compute capacity to continue to serve your traffic in the event of the loss of an entire zone (for single-region instances) or an entire region (for multi-region instances).

Metric	Maximum for single-region instances	Maximum per region for multi-region instances
High priority total	65%	45%
24-hour smoothed aggregate	90%	90%

To help you stay below the recommended maximums, create alerts in Stackdriver Monitoring (/spanner/docs/monitoring-stackdriver#create-alert) that track high-priority CPU utilization and the average CPU utilization over 24 hours.

CPU utilization can have an impact on request latencies. Overloading of an individual backend server will trigger higher request latencies. Applications should run benchmarks and active monitoring to verify that Cloud Spanner meets their performance requirements.

Thus, for performance-sensitive applications, you may need to further reduce CPU utilization using techniques described in the following section.

This section explains how to reduce an instance's CPU utilization.

In general, we recommend that you add nodes to your instance as a starting point. After you add nodes, you can investigate and address the root causes of high CPU utilization.

If you exceed the recommended maximums for CPU utilization, we strongly recommend adding nodes to your instance so it can continue to operate effectively. If you want to automate this process, you can create an application that monitors CPU utilization, then adds and removes nodes as needed, using the `UpdateInstance` (/spanner/docs/reference/rpc/google.spanner.admin.instance.v1#google.spanner.admin.instance.v1.InstanceAdmin.UpdateInstance) method.

To determine how many nodes you need, consider the peak high-priority CPU utilization as well as the 24-hour smoothed average. Always allocate enough nodes to keep the CPU utilization below the recommended maximums. As previously described, you may need to allocate extra nodes for performance-sensitive applications (for example, to accommodate workload spikes).

If you do not have enough nodes, Cloud Spanner postpones tasks by priority level. Low-priority system tasks, like database compaction and schema change validation, can be deferred in favor of user tasks. However, these tasks are critical to the health of your instance, and Cloud Spanner cannot defer them indefinitely. If Cloud Spanner cannot complete its low-priority system tasks within a certain time window—on the order of several hours to a day—due to

insufficient compute resources, Cloud Spanner might increase the priority of the system tasks. **This change affects the performance of user tasks**.

In some cases, your instance might have high CPU utilization because of SQL queries that are not as efficient as they could be. You can use the query statistics (/spanner/docs/query-statistics) for your database to identify queries that result in high CPU usage. Then, based on their query plans (/spanner/docs/query-execution-plans), you can optimize these queries to reduce CPU usage.

- Monitor your instance with the Cloud Console (/spanner/docs/monitoring-console) or the Stackdriver Monitoring console (/spanner/docs/monitoring-stackdriver).
- Create alerts for Cloud Spanner CPU utilization (/spanner/docs/monitoring-stackdriver#create-alert).
- Find out how to change the number of nodes (/spanner/docs/create-manage-instances#change-nodes) in a Cloud Spanner instance.
- Learn how to find correlations between high latency and other metrics (/spanner/docs/monitoring-stackdriver#create-charts).